Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

MAESTRÍA EN ECONOMÍA

TRABAJO FINAL DE MAESTRÍA

Estimación de la desigualdad de oportunidades en Argentina mediante técnicas de machine learning

AUTOR: ESTEBAN EMILIO RODRÍGUEZ

DIRECTOR: JORGE CARRERA

MAYO DE 2019

Resumen

La distribución de ingresos en una sociedad suele no depender sólo de esfuerzos y decisiones individuales. Cuando esto ocurre, se suele hablar de la existencia de desigualdad de oportunidades. El estudio tanto de su nivel como de sus determinantes es relevante por varios motivos. En primer lugar, la literatura sugiere que la desigualdad de oportunidades puede estar asociada al crecimiento económico en un grado mayor a la desigualdad de ingresos. En segundo lugar, su estudio permite una mejor comprensión sobre los mecanismos económicos e institucionales que generan la desigualdad de ingresos. Finalmente, existe evidencia respecto a que la sociedad apoya en mayor medida las políticas redistributivas cuando comprende que las desigualdades se originan en circunstancias exógenas al individuo. En la literatura se encuentran tanto estimaciones paramétricas como no paramétricas del nivel de desigualdad de oportunidades, las cuales han sido cuestionadas por presentar sesgos y ser dependientes de la modelización escogida. En este trabajo se realiza una adaptación de la metodología empleada por Brunori, Hufe y Mahler (2018) basada en árboles de regresión condicionales y bosques aleatorios condicionales, algoritmos que suelen incluirse en el amplio espectro del Machine Learning (ML) o Aprendizaje Automático. Según estos autores, con esta metodología se consiguen estimaciones más precisas, permitiendo considerar complejas interacciones entre las variables, a la vez que se incorporan nociones de significancia estadística que permiten el testeo de la existencia de igualdad de oportunidades. Utilizando datos de la Encuesta Permanente de Hogares (EPH) para Argentina en el período 2016-2018, se comprueba que esta nueva metodología produce mejores estimaciones de la distribución de los ingresos en nuestro país. Adicionalmente, se encuentra evidencia de la existencia de desigualdad de oportunidades con un 1% de significatividad tanto a nivel país como regional. Las variables más relevantes a la hora de explicar la desigualdad de oportunidades son género, el nivel educativo de los padres y el tamaño del hogar, aunque hay evidencia de que tanto los niveles de desigualdad como sus principales determinantes varían de acuerdo a la región del país. Los resultados obtenidos permiten una mejor comprensión sobre la desigualdad económica al interior de nuestro país, constituyendo un aporte tanto para el diseño de políticas públicas como para el mejoramiento de las estadísticas tendientes a evaluar las condiciones de vida de una sociedad. Adicionalmente, este trabajo contribuye a demostrar la utilidad de ciertos algoritmos de machine learning en el estudio de problemas económicos relevantes.

Palabras clave: desigualdad de oportunidades, distribución del ingreso, machine learning,
Código JEL: C53, D31, D63

Índice

1. Introducción

Existen distintas formas de medir qué tan desigual es una sociedad, no existiendo un consenso respecto a cuál es el tipo y grado de igualdad al cual se debería aspirar. Más aún, en materia de distribución de ingresos o de riqueza, unas sociedades parecen ser más tolerantes que otras en cuanto a los niveles de desigualdad deseables. De todas formas, en las sociedades democráticas modernas, se suele aceptar que todas las personas deberían gozar de las mismas oportunidades para obtener determinado nivel de ingresos, originando un interés por conocer cómo se distribuyen estas oportunidades. Si bien, en un primer momento, el estudio de la igualdad o desigualdad de oportunidades estuvo circunscripto al terreno de la filosofía política, siendo la Teoría de la Justicia de John Rawls (1971) el más claro ejemplo de esto, el interés de la teoría económica por este tema ha ido creciendo.

1.1 La igualdad de oportunidades como preocupación económica

Los motivos por el creciente interés de los economistas en esta problemática son varios. En primer lugar, se ha argumentado que la desigualdad de oportunidades está asociada al crecimiento económico en un grado aún mayor a la desigualdad de ingresos. Marrero y Rodríguez (2010) analizaron la desigualdad de ingresos de Estados Unidos, separando entre desigualdad de oportunidades y la que llaman 'desigualdad de retornos al esfuerzo'. Hallaron evidencia robusta de una relación positiva entre esta última y el crecimiento, y de una relación negativa entre desigualdad de oportunidades y crecimiento económico. Estos resultados compatibilizan la visión de que cierta desigualdad de ingresos es positiva para los incentivos que movilizan una economía capitalista, con la visión de que ciertas personas no consiguen alcanzar todo su potencial productivo por motivos ajenos a su voluntad. En este sentido, el Banco Mundial (2006) sostiene que la existencia de fuertes y persistentes desigualdades en las oportunidades iniciales en los individuos genera trampas de desigualdad que representan severas restricciones al crecimiento económico.

En segundo lugar, Checchi y Peragine (2010) sostienen que el estudio de la desigualdad de oportunidades permite una mejor comprensión sobre los mecanismos económicos e institucionales que generan la desigualdad de ingresos. En particular, existen interdependencias entre la desigualdad de ingresos y la de oportunidades, que suelen visualizarse en un reducido grado de movilidad social: los niños de hogares pobres suelen ser pobres en su adultez y lo contrario ocurre con quienes nacen en hogares ricos. Corak (2012) realiza estimaciones de la 'elasticidad intergeneracional de los ingresos', encontrando

que en países avanzados como EEUU y Gran Bretaña el ingreso de una persona está muy correlacionado con el de sus padres, indicando que el esfuerzo y talento individual no consiguen superar las diferencias de nacimiento. Por otro lado, en Canadá y en los países nórdicos la mayor movilidad social es significativamente mayor. De acuerdo con el autor, estas diferencias entre países tienen que ver con el rol de tres instituciones: la familia, el mercado laboral y el Estado. Cuanto mayor es la capacidad de las familias para invertir en sus hijos, tanto en términos monetarios como no monetarios, mayores las oportunidades que tendrán los chicos en su vida. Cuanto mayor es la igualdad en los retornos a la educación en el mercado laboral, mayores los incentivos para invertir en educación. Si el Estado implementa políticas públicas progresivas, la relevancia del entorno familiar se reduce y aumenta la igualdad de oportunidades.

Relacionado con el posible rol del Estado, existe evidencia respecto a que la sociedad apoya en mayor medida las políticas redistributivas cuando percibe que las desigualdades tienen un origen injusto (Alesina y La Ferrara, 2005). Más aún, Jiménez (2016) sostiene que la desigualdad económica percibida como desigualdad de oportunidades es, probablemente, una de las principales fuentes de descontento e inestabilidad social y política. En este sentido, Ferreira y Gignoux (2011) afirman que existe una creciente visión normativa según la cual las políticas públicas deberían orientarse hacia la búsqueda de la igualdad de oportunidades y no a la igualdad de ingresos o riqueza. Más allá de que este punto es opinable, contribuye a explicar el creciente interés en la desigualdad de oportunidades.

Sin embargo, más allá de la relevancia que se le asigna al tema, no existe un consenso ni a nivel filosófico ni en los debates públicos respecto a cómo debe definirse la igualdad de oportunidades (Lefranc, Pistolesi y Trannoy, 2009). Del mismo modo, tampoco hay una forma estandarizada de medir el nivel de desigualdad de oportunidades.

Dado el carácter empírico de este trabajo, se adoptará una definición de igualdad de oportunidades similar a la utilizada en gran parte de la literatura previa y que puede ser estimada con los datos disponibles. Para justificar la utilización de los algoritmos que aquí se emplean, se hará una comparación del poder predictivo de las principales metodologías de estimación empleadas en la literatura con el de los árboles de regresión y bosques aleatorios condicionales desarrollados por Hothorn, Hornik, y Zeileis (2006). Una vez validados estos métodos, se los utilizará para encontrar evidencia de la existencia de desigualdad de oportunidades en Argentina, estimar su nivel e identificar sus principales determinantes. Al hacer esto, se espera identificar diferencias entre distintas regiones del

país respecto al nivel y determinantes de la desigualdad de oportunidades, como también encontrar interacciones entre las variables que, a priori, no son planteadas en la modelización. La única fuente de información es la Encuesta Permanente de Hogares (EPH) elaborada por el Instituto Nacional de Estadísticas y Censos (INDEC), considerando el período 2016-2018, dado que en el mismo no se han producido cambios metodológicos relevantes ni en el diseño de la encuesta ni en la recolección de los datos.

1.2 Estructura y motivación del trabajo

La estructura del trabajo es la siguiente. En el Capítulo 2 se formulan las principales preguntas de investigación, junto con los objetivos y las hipótesis que guían este trabajo. La revisión del marco teórico se dividirá en dos capítulos consecutivos. En el Capítulo 3, se repasará la literatura relevante sobre el tema de la igualdad de oportunidades, discutiendo las distintas definiciones junto con una revisión crítica de las estrategias de estimación empírica del nivel de desigualdad de oportunidades. En el Capítulo 4, luego de discutir qué se entiende por Machine Learning (ML), se realiza una revisión de los algoritmos basados en árboles, desde los árboles de regresión introducidos por Breiman, Friedman, Olshen, y Stone (1984) hasta los bosques aleatorios, tal vez el algoritmo de ML más difundido hoy en día. En particular, se explicarán los árboles de regresión condicionales y los bosques aleatorios condicionales, dado que son los dos métodos que se utilizan en este trabajo. En el Capítulo 5 se presentarán algunos detalles metodológicos, explicando qué variables de la EPH serán utilizadas. Los resultados de la investigación se presentan en el Capítulo 6, donde se espera mostrar evidencia a favor de todas las hipótesis planteadas en el Capítulo 2. En el Capítulo 7 se discuten las principales conclusiones del trabajo, junto con una serie de propuestas para investigaciones futuras. Luego de las referencias bibliográficas, el trabajo finaliza con un breve anexo con consideraciones sobre el software utilizado e indicaciones para consultar el anexo online, donde están disponibles resultados adicionales no incluidos en este documento, como también el código fuente que permite la reproducción de los resultados.

Respecto a la motivación de este trabajo, sobresale la posibilidad de combinar diferentes intereses del autor de una manera coherente y, en cierta forma, novedosa. La problemática de la desigualdad es un tema que ya ha sido trabajado de manera conjunta por el autor y el tutor de esta tesis (Carrera, Rodríguez y Sardi, 2016), quienes incluso ya habían abordado los vínculos de la desigualdad de oportunidades con la desigualdad de ingresos y sus posibles impactos macroeconómicos (Carrera y Rodríguez, 2013). Por otro lado, en los

últimos tiempos el autor ha estado siguiendo con interés el impulso que están tomando los algoritmos de ML para distinto tipo de aplicaciones. La posibilidad de combinar ambas inquietudes en un mismo trabajo era una tarea que estaba pendiente.

Antes de continuar con el siguiente capítulo, una aclaración respecto al vocabulario utilizado en este trabajo. La bibliografía sobre ML ha crecido exponencialmente en los últimos años, estando predominantemente en inglés. Como resultado, gran parte de su terminología no tiene una traducción al español estandarizada y no se pretende imponerla aquí. Más aún, se ha considerado que la utilización de traducciones tales como 'aprendizaje automático' podría confundir en lugar de ayudar al lector familiarizado con ML. Por lo tanto, salvo cuando las traducciones sean obvias, como en el caso de los mencionados árboles de regresión y bosques aleatorios[1], se utilizarán los términos en inglés.

2. Pregunta de investigación, objetivos e hipótesis de trabajo

A la hora de analizar la distribución del ingreso en la Argentina, ¿existe la igualdad de oportunidades? Responder a este interrogante requiere responder un par de preguntas previas. En primer lugar, ¿qué es la igualdad de oportunidades?, La literatura suele desagregar el origen de los resultados individuales -en este caso, los ingresos- en dos tipos de factores: aquellos que constituyen una fuente legítima de desigualdad, denominados 'esfuerzos', y aquellos donde las desigualdades que producen son social o moralmente inaceptables, denominados 'circunstancias' (Roemer, 1998). A partir de allí, una definición posible es considerar que existe la igualdad de oportunidades cuando la distribución de ingresos es independiente de las circunstancias que enfrentan distintos grupos de la población. Si bien la discusión respecto a qué se incluye como circunstancia es en gran parte subjetiva, la selección de las mismas suele estar limitada por cuestiones metodológicas o de disponibilidad de datos. En este sentido, la segunda pregunta que hay que responder es la siguiente: ¿cuál es la mejor manera de estimar la desigualdad de oportunidades?

La mayor parte de la literatura empírica se ha visto forzada a considerar un número reducido de circunstancias, dado que ninguna muestra es lo suficientemente grande para las metodologías usuales. Por un lado, las estimaciones paramétricas pueden estimar con precisión sólo un número reducido de parámetros. Por el otro, las estimaciones no paramétricas enfrentan el problema de que no pueden clasificar a la muestra en un gran

[1] Traducción de regression trees y random forest.

número de grupos, ya que algunos quedarían con muy pocas observaciones como para realizar una estimación precisa. De esta forma, las estimaciones de la desigualdad de oportunidades bajo ambas metodologías resultan sesgadas (Brunori, Peragine, y Serlenga, 2016) y con modelizaciones extremadamente simples.

Considerar un número amplio de circunstancias admitiendo complejas interacciones entre las mismas requiere implementar metodologías donde el tamaño de la muestra no sea un limitante para la cantidad de variables a considerar. Muchos algoritmos de ML cumplen con este requisito, por lo que permitirían, en principio, tanto una estimación más precisa de la desigualdad de oportunidades como también una mejor identificación de sus principales determinantes. En este trabajo se comparará el desempeño de los árboles de regresión condicionales y bosques aleatorios condicionales desarrollados por Hothorn et al. (2006) con el de las metodologías tradicionales. Después de validar estos algoritmos como metodología para estudiar la desigualdad de oportunidades, se los utilizará para estimar el nivel y los determinantes de la desigualdad de oportunidades en Argentina.

De esta forma, el objetivo general de este trabajo puede plantearse de la siguiente manera:

2.1 Objetivo general: Estudiar si existe desigualdad de oportunidades en la Argentina, identificando sus principales determinantes.

Avanzando en el detalle de lo que se intenta realizar en este trabajo, se plantean los siguientes objetivos específicos:

2.2 Objetivos específicos:

1. Analizar si los algoritmos de ML mejoran las estimaciones de los niveles de ingresos respecto a las técnicas econométricas tradicionalmente encontradas en la literatura.
2. Buscar evidencia robusta de que, en el contexto de la distribución de los ingresos, existe desigualdad de oportunidades en Argentina
3. Identificar los principales determinantes de la desigualdad de oportunidades en Argentina.
4. Determinar si hay diferencias entre las distintas regiones del país en cuanto al nivel de desigualdad de oportunidades y sus determinantes.

Utilizando datos de la EPH, el análisis se restringirá a los diez trimestres comprendidos entre el segundo trimestre de 2016 y el tercero de 2018, dado que no se produjeron cambios metodológicos de relevancia en ese período. Además de comparar la calidad de las estimaciones que produce cada metodología, se realizarán estimaciones del nivel de desigualdad de oportunidades tanto a nivel país como al interior de cada una de las regiones identificadas por la EPH (Gran Buenos Aires, Noroeste, Nordeste, Cuyo, Pampeana y Patagonia). También se hará un análisis de los principales determinantes de la desigualdad de oportunidades y de las diferencias observadas entre las distintas regiones del país.

En los objetivos que se acaban de plantear, implícitamente se está asumiendo que, en materia de distribución de los ingresos, existe desigualdad de oportunidades en la Argentina. En efecto, esta es la hipótesis principal de este trabajo.

2.3 Hipótesis Principal: En Argentina, el ingreso personal está condicionado por factores que están fuera del control individual (circunstancias). Es decir, existe desigualdad de oportunidades.

Más aún, en base a las *circunstancias* que pueden ser consideradas por la EPH y a resultados sugeridos por la literatura previa, pueden arriesgarse las siguientes hipótesis secundarias.

2.4 Hipótesis Secundarias:

1. Los algoritmos de ML permiten estimaciones más precisas del nivel de desigualdad de oportunidades que enfrenta una sociedad, comparando con las estimaciones paramétricas y no paramétricas frecuentemente utilizadas en la literatura.
2. El nivel de ingresos de una persona está condicionado por el nivel educativo de sus padres.
3. Las mujeres obtienen, en promedio, menor nivel de ingresos que los hombres.
4. Los determinantes de la desigualdad de oportunidades en Argentina varían de una región a otra del país.

En los próximos dos capítulos se discute el marco teórico relevante para este trabajo, mientras que en el capítulo 5 se presentan los detalles específicos sobre la selección de las variables para el trabajo empírico. La evidencia que intenta demostrar las hipótesis planteadas será discutida en el Capítulo 6.

3. Aspectos teóricos y empíricos de la desigualdad de oportunidades

A la hora de analizar la distribución de ingresos de una sociedad suelen plantearse preguntas respecto a qué tan igualitaria es la misma, dando lugar a la utilización de indicadores como el coeficiente de Gini, el de Theil, los *Top Incomes*, entre muchos otros. Sin embargo, salvo en casos muy extremos, es difícil establecer si determinado nivel de desigualdad de ingresos es un problema que debe ser resuelto por algún tipo de intervención política o si, por el contrario, contribuye a generar los incentivos adecuados para el desenvolvimiento de una economía capitalista. Si bien en el debate público suelen surgir posiciones extremas -toda desigualdad es mala o toda desigualdad es buena-, desde la filosofía política se planteó la necesidad de distinguir entre distintos tipos de desigualdades. John Rawls, en su Teoría de la Justicia (Rawls, 1971) sostiene la necesidad de distinguir entre la desigualdad social o moralmente justificada, de aquella injustificada. La adaptación que realizó la literatura económica a esta distinción consiste en la separación entre 'desigualdad de resultados' y 'desigualdad de oportunidades', ya sea para oponerse a cualquier tipo de política redistributiva (Friedman y Friedman, 1979) como para refinar el tipo de igualdad al que debería aspirarse sin violar las libertades individuales (Sen, 1985). Más allá de los distintos enfoques que se le pueden dar a esta cuestión, existe un mayor consenso respecto a que todas las personas deberían tener las mismas oportunidades de obtener determinado nivel de ingresos, respecto a la propuesta de igualar los ingresos *ex - post*. Sin embargo, a la hora de definir qué se entiende por igualdad de oportunidades, el consenso está ausente nuevamente.

En la sección primera de este capítulo se abordará la discusión sobre cómo definir el concepto de igualdad de oportunidades, explicando la definición adoptada en este trabajo. La segunda sección se centra en un aspecto particular de esta discusión, la de definir cuáles son aquellos factores que están fuera del control individual y que dan origen a la desigualdad de oportunidades. En la sección tercera se presentan de manera estilizada las distintas metodologías empleadas en la literatura para medir la desigualdad de oportunidades, haciendo énfasis en los inconvenientes que presentan y que justifican la búsqueda de una metodología superadora. Por último, en la sección cuarta se introduce brevemente la metodología sugerida basada en árboles y bosques de regresión, tema sobre el que se profundizará en el Capítulo 4.

3.1 ¿Qué es la igualdad de oportunidades?

El punto de partida para cualquier definición de igualdad de oportunidades suele ser la idea de que los resultados individuales tienen su origen en dos tipos de factores: aquellos que constituyen una fuente legítima de desigualdad, denominados esfuerzos, y aquellos donde las desigualdades que producen son social o moralmente inaceptables, denominados circunstancias. Qué factores deben considerarse esfuerzos o circunstancias es motivo de debate y entra en el terreno de la teoría normativa. Dejaremos esta discusión para la siguiente sección. Por el momento, nos podemos quedar con la conclusión de Lefranc et al. (2009), quien afirma que, por lo menos, el conjunto de circunstancias debe incluir el trasfondo social del individuo. Es decir, de estar disponible la información, se deben considerar factores como raza, género, lugar de nacimiento, familia y grupo socioeconómico, entre otros.

Suponiendo que estamos de acuerdo en el conjunto de circunstancias a considerar, existen dos caminos distintos para definir la igualdad de oportunidades. Paes de Barros, Ferreira, Molinas Vega y Saavedra Chanduvi (2009) identifican, en primer lugar, una visión meritocrática bajo la cual personas con idénticas elecciones y nivel de esfuerzo deben obtener idénticos resultados. En segundo lugar, mencionan una visión que los autores definen como igualitaria, la cual sostiene que la distribución de resultados tiene que ser estocásticamente independiente de las circunstancias. Esta segunda visión es más exigente, ya que requiere que las circunstancias que enfrenta cada persona no condicionen los esfuerzos que puede realizar.

Más formalmente, siguiendo la notación de Brunori et al (2018), supongamos una población de N individuos, cada uno con un ingreso y_i. Este ingreso individual es resultado de dos grupos de factores. En primer lugar, un conjunto de P circunstancias que están fuera del control del individuo: $\Omega_i = \{C_i^1, \dots, C_i^p, \dots, C_i^P\}$. En segundo lugar, un conjunto de Q esfuerzos $\Theta_i = \{E_i^1, \dots, E_i^q, \dots, E_i^Q\}$. El ingreso individual es el resultado de una función $g: \Omega \times \Theta \to \mathbb{R}_+$, que puede ser escrita como $y_i = g(\Omega_i, \Theta_i)$.

Cada circunstancia $C^p \in \Omega$ tiene un total de X^p realizaciones posibles, donde cada realización para el individuo i se denomina x_i^p. De acuerdo a estas realizaciones de las circunstancias, se definen los 'tipos' $T = \{t_1, \dots, t_m, \dots, t_M\}$, cada uno de los cuales agrupa a un sector de la población uniforme en término de circunstancias. Es decir, los individuos $i\ y\ j$ pertenecen al mismo tipo $t_m \in T\ si\ x_i^p = x_j^p\ \forall\ C^p \in \Omega$. Del mismo modo, los individuos $i\ y\ j$ pertenecen a distintos tipos si $\exists\ C^p \in \Omega :\ x_i^p \neq x_j^p$.

Con la notación anterior, Brunori et al (2018) expresan las dos definiciones de igualdad de oportunidades dadas por Roemer (1998). Por un lado, una definición *ex – ante*, que se enfoca en las diferencias entre tipos, sin tener en cuenta los esfuerzos. Desde esta perspectiva, la igualdad de oportunidades se cumple cuando el resultado o ingreso promedio de cada tipo es igual al promedio poblacional. Es decir, cuando la distribución del ingreso es independiente del tipo al que se pertenece. Por otro lado, existe una definición *ex – post*, que se enfoca en el resultado individual condicional al esfuerzo realizado. De acuerdo a esta definición, la igualdad de oportunidades existe cuando individuos con igual nivel de esfuerzos obtienen los mismos resultados. La definición *ex – ante*, puede asociarse a la visión igualitaria de Paes de Barros et al. (2009), mientras que la definición *ex – post* se corresponde con la visión meritocrática.

Definición de Igualdad de Oportunidades ex – ante (visión igualitaria)

$$F(y_i|\Omega_i) = F(y_i)$$

$$\forall\, t_i, t_j \in T,\ \bar{y}_{t_i} = \bar{y}_{t_j}$$

Definición Igualdad de Oportunidades ex – post (visión meritocrática)

$$\Theta_i = \Theta_j \Rightarrow \mathrm{y}_i = \mathrm{y}_j$$

Cabe mencionar que, si la distribución de los esfuerzos es independiente del tipo, la igualdad de oportunidades ex – post implica la igualdad de oportunidades ex – ante (Ramos y Van de Gaer, 2012). Pero esta independencia suele ser difícil de justificar, por lo que, en general, estas dos definiciones expresan visiones distintas y conflictivas de lo que se considera igualdad de oportunidades (Checchi, Peragine y Serlenga, 2015).

En este trabajo se adoptará la definición ex – ante o igualitaria de la igualdad de oportunidades, por la sencilla razón de que se disponen datos para las circunstancias que enfrentan los individuos, o al menos para algunas de ellas, siendo mucho más difícil de cuantificar los esfuerzos individuales. En la sección siguiente retomamos el debate sobre qué variables pueden considerarse dentro del grupo de circunstancias, mientras que en la sección 2.3 se explicarán las principales metodologías utilizadas para medir la desigualdad de oportunidades ex – ante.

3.2 La delimitación del conjunto de circunstancias.

Para diversos autores, a la hora de determinar qué se considera una circunstancia el criterio relevante debe ser la responsabilidad individual. Por ejemplo, Checchi y Peragine (2010) citan los llamados Principios de Compensación y de Responsabilidad. El primero de

estos principios establece que las diferencias en los resultados individuales atribuidas a factores fuera del control individual son inequitativas y deben ser compensadas por la sociedad. Por otro lado, el segundo principio establece que diferencias en los resultados explicadas por factores sobre los cuales el individuo es responsable, son equitativas y no deben ser compensadas. Aun así, Roemer (1998) resalta que esta delimitación es estrictamente política y que, en distintas sociedades, puede variar la percepción que se tiene sobre si determinada variable está o no bajo control del individuo. Estas distintas percepciones sociales son ejemplificadas por Lefranc et al. (2009), quien sostiene que, en Estados Unidos, el éxito individual se suele atribuir mayormente al esfuerzo individual, mientras que en sociedades europeas se enfatiza el rol de la suerte. En efecto, la consideración de factores aleatorios o "suerte" complica aún más la delimitación de las circunstancias.

En el contexto de la desigualdad de ingresos, Vallentyne (2002) distingue dos tipos de suerte. Por un lado, una suerte inicial, definida como el conjunto de factores que tienen influencia en los resultados del individuo hasta el momento en que el mismo es responsable de sus decisiones. Más allá de ese momento, existe una suerte tardía, que sigue afectando los resultados del individuo. Dentro del primer tipo de suerte, se encuentra todo el trasfondo social en el que nace y se educa el individuo, aunque también incluye cierta suerte genética, en el sentido que la persona puede haber heredado algún talento en particular. Como señala Lefranc et al. (2009), la mayoría de los autores coincide en que las desigualdades originadas en el contexto social deben compensarse -son injustas-, pero que la suerte genética no debe compensarse porque se entraría en conflicto con otros valores éticos. Menos claro es qué debe hacerse con la suerte tardía, como puede ser ganarse la lotería o salir sorteado para ir a la guerra, ya que este tipo de suerte puede no ser completamente aleatoria, sino estar vinculada tanto a esfuerzos como a circunstancias previas.

El grado de subjetividad en la selección de las circunstancias es explicitado por Ramos y Van de Gaer (2012), quienes sostienen que existen al menos tres visiones dentro de la filosofía política respecto a la cuestión de qué variables son responsabilidad de los individuos. Una primera visión sostiene que un individuo es responsable de lo que está bajo su control, donde control está asociado al reconocimiento de la existencia de la libre voluntad. Quienes niegan la existencia de la libre voluntad, incluyen todas las variables observables dentro del conjunto de circunstancias, de tal modo que toda desigualdad es considerada injusta. Por otro lado, quienes aceptan la existencia de la libre voluntad, clasifican como circunstancias a variables del entorno familiar como educación y ocupación

de los padres, junto con características individuales como género, raza, edad y coeficiente intelectual. Bajo esta visión, variables de contexto como el acceso a servicios básicos como agua, sanitación, electricidad o transporte, también deberían incluirse dentro del conjunto de circunstancias.

Una segunda visión sostiene que los individuos son responsables de sus preferencias y de las elecciones que toman en base a las mismas. Bajo esta visión, el conjunto de circunstancias puede reducirse a un mínimo, ya que si se considera, por ejemplo, que las preferencias varían de acuerdo a variables como sexo o raza, las diferencias observadas en los resultados entran en el terreno de la responsabilidad individual y no en un tratamiento discriminatorio. Una tercera visión, vinculada a la idea de que uno es propietario de sí mismo y de todas las características personales, conduce al extremo de considerar vacío al conjunto de circunstancias. De esta forma, no hay lugar para la desigualdad de oportunidades, siendo todas las desigualdades observadas legítimas.

En este trabajo se adopta un enfoque más cercano a la primera de estas visiones, incluyendo dentro de las circunstancias tanto a variables del entorno familiar como a ciertas características individuales. Cabe mencionar que Gasparini (2002) adopta un enfoque conceptualmente distinto, distinguiendo entre fuentes de desigualdad socialmente aceptadas y no aceptadas, siendo el enfoque seguido por Serio (2011). De todas formas, esta clasificación de las variables sigue siendo igualmente subjetiva y, a la hora del trabajo empírico, conduce aproximadamente a la misma selección de las variables. En este sentido, en la Tabla 1, se presenta un resumen de las circunstancias que han sido consideradas en la literatura más cercanas a este trabajo.

De todas formas, la selección de las circunstancias en la literatura empírica no depende exclusivamente de argumentos teóricos. En primer lugar, depende de la disponibilidad de datos. En el caso particular de este trabajo, hubiera sido deseable disponer de mayor información sobre el entorno en que se criaron los individuos que hoy obtienen ingresos, pero la información de este tipo que puede obtenerse de la EPH es muy limitada. En segundo lugar, aun disponiendo de información sobre determinadas circunstancias, la metodología escogida puede verse restringida por la cantidad de observaciones, obligando a descartar algunas circunstancias, trabajando solamente con un subconjunto de las mismas. Este tema será abordado en la siguiente sección, donde se repasan de manera estilizada las principales metodologías de estimación de la desigualdad de oportunidades ex – ante.

Tabla 1. Circunstancias consideradas en la literatura

Estudio	Circunstancias vinculadas al entorno familiar	Circunstancias vinculadas a características individuales
Bourguignon, Ferreira y Menéndez (2007)	Educación del padre y de la madre; ocupación del padre.	Raza; nacionalidad.
Paes de Barros et al. (2009)	Área de residencia; género del jefe de hogar; nivel de educación de los padres; ingreso familiar per cápita; cantidad de hermanos; presencia de los dos padres en el hogar.	Género
Lefranc et al. (2009)	Grupo ocupacional del jefe de hogar.	
Checchi y Peragine (2010)	Lugar de residencia; nivel educativo del padre.	Género
Serio (2011)	Educación de los padres; ocupación de los padres	Género, lugar de nacimiento.
Checchi et al. (2015)	Educación de los padres; categoría ocupacional de los padres.	Género; nacionalidad; grupo de edad.
Brunori et al. (2016)	Lugar de nacimiento; ocupación del padre y de la madre; educación del padre y de la madre.	Género
Brunori et al. (2018)	Lugar de nacimiento; presencia de padres en el hogar; cantidad de adultos en el hogar; cantidad de adultos que trabajan en el hogar; cantidad de niños en el hogar; país de nacimiento del padre/madre; nivel educativo del padre/madre; status ocupacional del padre/madre; principal ocupacional del padre /madre; régimen de tenencia de la vivienda.	Género

3.3 ¿Cómo se mide la desigualdad de oportunidades?

En la literatura se encuentran distintas formas de estimar el nivel de desigualdad de oportunidades ex - ante. Sin embargo, más allá de las especificidades de cada trabajo, podemos resumir el proceso de la siguiente manera. Una vez que se seleccionaron las circunstancias, se estima la distribución de los ingresos generada por estos factores que están fuera del control individual. Esta distribución de ingresos estimada, denominada contrafactual o suavizada (Ferreira y Gignoux, 2011). Al estar explicada por las circunstancias, esta distribución contrafactual es una distribución de ingresos 'injusta', es decir, producto de la existencia de desigualdad de oportunidades. El paso siguiente consiste en aplicar un indicador de desigualdad a esta distribución de ingresos contrafactual, por ejemplo, el coeficiente de Gini, dando lugar a una estimación del nivel de desigualdad de oportunidades.

Este procedimiento implica que la calidad de la estimación del nivel de desigualdad de oportunidades depende directamente de la calidad de la estimación de la distribución de ingresos contrafactual. De esta forma, al existir distintas alternativas para realizar esta estimación, es necesario analizar cuál es la más apropiada. En particular, la literatura analizada puede clasificarse de acuerdo a tres metodologías de estimación: paramétrica, no paramétrica y una tercera metodología más reciente, basada en algoritmos de ML, que es la elegida en este trabajo. A continuación, se describen las metodologías paramétrica y no paramétrica. De la tercer metodología sólo se realiza una introducción, destinando el Capítulo 4 a una discusión más profunda sobre la misma.

3.3.1 Estimaciones No Paramétricas

Esta fue la metodología empleada en el trabajo seminal de Van de Gaer (1995), la cual fue seguida por, entre otros, Checchi y Peragine (2010) y Checchi et al. (2015). Consiste en los siguientes pasos:

1. Se selecciona el set de circunstancias $\Omega_i = \{C_i^1, \dots, C_i^p, \dots, C_i^P\}$ y las X^p realizaciones permitidas para cada una de ellas.
2. Se construyen los tipos $T = \{t_1, \dots, t_m, \dots, t_M\}$ en base a todas las combinaciones posibles de circunstancias. De esta forma, la muestra queda dividida en *M* tipos disjuntos, cada uno de los cuales alberga individuos con exactamente las mismas circunstancias, mientras que individuos de distintos tipos presentan diferencias en al menos una de las circunstancias.
3. Para cada individuo $i \in t_m$, se estima un nivel de ingresos y_i^E igual al ingreso promedio al interior de su tipo:

$$y_i^E = \frac{1}{N_m} \sum_{j \in t_m} y_j$$

4. Se repite lo anterior para todos los tipos, asignando a cada uno de los *N* individuos el ingreso promedio de su tipo. De esta forma se obtiene la distribución de ingresos contrafactual $Y^E = \{y_1^E, \dots, y_i^E, \dots, y_N^E\}$.
5. Sobre esta distribución de ingresos contrafactual se aplica un indicador de desigualdad $I(\cdot)$, como podría ser el coeficiente Gini.

Para ejemplificar esta metodología, supongamos que, para una población conformada por 100 individuos, disponemos de información de los ingresos y de las

circunstancias “Género” y “Raza”, donde cada una de éstas puede tomar dos valores posibles:

$$C^1: Género = \{Varón, Mujer\}$$

$$C^2: Raza = \{Blanca, Otra\}$$

De esta forma, la población se divide en 4 tipos:

$$T = \{\{Varón;\ Blanca\}, \{Varón;\ Otra\}, \{Mujer;\ Blanca\}, \{Mujer;\ Otra\}\}$$

Supongamos que cada uno de estos tipos está conformado por 25 individuos, siendo los ingresos promedios de $1.000, $700, $800 y $600 respectivamente. De esta forma, se puede construir la distribución de ingresos contrafactual para los 100 individuos de la población:

$$Y^E = \{1.000, \dots, 1.000, 700, \dots, 700, 800, \dots, 800, 600, \dots, 600\}$$

Esta distribución de ingresos está explicada sólo por las circunstancias, por lo que puede considerarse injusta. De esta forma, aplicando un indicador de desigualdad a la misma, se obtiene una estimación del nivel de desigualdad de oportunidades que enfrenta esta población.

Entre las ventajas de esta metodología sobresale que es capaz de captar interacciones entre las variables sin forzar la forma en que las variables afectan el ingreso. Es decir, no es necesario forzar una influencia lineal y aditiva, por citar la modelización más común. Pero presenta un gran inconveniente y es que, dado el tamaño de la muestra, la cantidad de circunstancias a considerar está limitada por la cantidad de observaciones que quedan asignados a cada tipo. Por ejemplo, supongamos que en el ejemplo anterior queremos considerar una tercer circunstancia, la cual puede tomar dos valores:

$$C^3: Instrucción\ del\ Padre = \{Sabe\ leer\ y\ escribir, Otra\}$$

Esto eleva la cantidad de tipos a $2^3 = 8$, dadas las distintas combinaciones que se pueden hacer de estas circunstancias[2]. Dada que la muestra consta de sólo 100 individuos, puede verse que la estimación del ingreso al interior de cada tipo se hará con un número muy reducido de observaciones. Más aún, este problema crece rápidamente a medida que se consideran más variables y se admiten más valores para cada una. Por ejemplo, Checchi y Peragine (2010) utilizan cinco circunstancias que pueden tomar entro dos y tres valores cada una, dando lugar a 108 tipos, algunos de los cuales cuentan con menos de cinco

[2] *T={ {Varón; Blanca; Sabe leer y escribir}, {Varón; Blanca; Otra}, {Varón; Otra; Sabe leer y escribir}, {Varón; Otra; Otra}, {Mujer; Blanca; Sabe leer y escribir}, {Mujer; Blanca; Otra}, {Mujer; Otra; Sabe leer y escribir}, {Mujer; Otra; Otra} }.*

observaciones. Checchi et al. (2015) también utilizan cinco circunstancias, alguna de las cuales puede tomar cuatro valores, dando lugar a 144 tipos.

Un inconveniente adicional que presentan las estimaciones no paramétricas es que, si se desea considerar una circunstancia que toma valores continuos, como podría ser el ingreso del padre, el investigador debe decidir cómo particionar esta variable: en percentiles, cuartiles, o de otra manera, decisión que condiciona los resultados. Esto sucede porque, por lo general, la cantidad de realizaciones únicas observadas en una variable continua suele ser muy grande como para construir un tipo de acuerdo a cada valor único, lo que obliga a agruparlas de alguna manera arbitraria. Por ejemplo, para conformar los tipos de acuerdo a la variable ingreso del padre, pueden considerarse dos realizaciones posibles: ingreso del padre menor a la mediana, o mayor o igual a la mediana. Pero existen muchas otras opciones, siendo difícil establecer cuál es la más apropiada.

Este tipo de inconvenientes no es exclusivo de la metodología descrita anteriormente. Lefranc et al. (2009) siguen una metodología distinta, realizando tests de dominancia estocástica para testear la existencia de desigualdad de oportunidades. Aun así, su trabajo sufre de la misma restricción de no poder utilizar un gran número de circunstancias. Los autores mencionan que cuentan con información del grupo ocupacional (6 valores posibles) del jefe de hogar, su esposa y los padres de ambos, lo que permitiría una detallada clasificación del origen social del hogar. Sin embargo, argumentan que el uso de todas estas variables daría lugar a submuestras muy pequeñas y estimaciones poco precisas, por lo que sólo utilizan el grupo ocupacional del jefe de hogar.

En resumen, el principal inconveniente del enfoque no paramétrico es que, de todas las circunstancias observables, el investigador debe o bien seleccionar un subconjunto de ellas, o limitar los valores que pueden tomar, o ambas cosas para poder obtener estimaciones confiables. Particionar la muestra en un número reducido de tipos implica asumir un rol muy limitado para las circunstancias y, en la mayoría de los casos, lleva a una subestimación de la desigualdad de oportunidades (Ferreira y Gignoux, 2011). Pero intentar resolver este inconveniente incluyendo todas las circunstancias observables genera estimaciones con mayor varianza, lo cual puede resultar en una sobreestimación del nivel de desigualdad de oportunidades (Brunori et al. 2016).

Cabe destacar que la metodología basada en árboles de regresión, sobre la cual se hablará en la sección 3.3.3 y, con más detalle, en el Capítulo 4, tiene muchas similitudes con el método de estimación no paramétrica. También se divide a la población en tipos de acuerdo a las circunstancias, calculándose el nivel de ingresos promedio al interior de cada

uno de ellos. Pero hay una diferencia fundamental: no se construyen todos los tipos posibles, sino sólo un subconjunto de los mismos. Esto permite considerar un número mucho mayor de circunstancias. Y lo que resulta igualmente relevante, la selección de qué tipos se utilizan y cuáles no, no queda a cargo del investigador, sino que la realiza el algoritmo en base a criterios estadísticos. Pero antes de profundizar con estos métodos, se discutirá la otra metodología de estimación ampliamente difundida en la literatura de igualdad de oportunidades.

3.3.2 Estimaciones Paramétricas

Una alternativa para la estimación de la distribución del ingreso contrafactual es mediante regresiones paramétricas. Dado que es, probablemente, la técnica econométrica más difundida, requiere una explicación menos detallada. Consiste básicamente en estimar el logaritmo del ingreso mediante una función lineal de las circunstancias consideradas:

$$\ln(y_i) = \beta_0 + \sum_{c=1}^{K} \beta_c C_i^c + \varepsilon_i$$

De esta forma, la distribución contrafactual se construye con los valores predichos por la regresión anterior:

$$y_i^E = e^{\widehat{\beta}_0 + \sum_{c=1}^{K} \widehat{\beta}_c C_i^c}$$

Es importante destacar que, tal como ocurría con la estimación no paramétrica, la regresión anterior va a predecir ingresos iguales para individuos con las mismas circunstancias. Es decir, si bien no se construyen formalmente los *tipos*, es como si se lo hiciera, por lo que tiene sentido seguir hablando de los mismos.

La estimación paramétrica puede realizarse mediante, por ejemplo, Mínimos Cuadrados Ordinarios, y suele brindar mejores resultados que los métodos no paramétricos cuando la muestra no es demasiado grande (Ferreira y Gignoux, 2011), permitiendo considerar un número mayor de circunstancias. Otra virtud de esta metodología es que permite una mejor identificación de la contribución individual de cada variable, algo no tan sencillo con las metodologías alternativas. Sin embargo, también tiene sus inconvenientes. En la versión más simple de la regresión anterior, se supone que todas las variables tienen un efecto lineal sobre los ingresos, ignorando todo tipo de interdependencias entre las variables. Por supuesto, se puede elevar el grado del polinomio o agregar interacciones entre las variables, pero esto daría lugar a un modelo cada vez más saturado, expuesto a los mismos problemas de sobreestimación de los métodos no paramétricos.

La estimación paramétrica es la metodología empleada por Bourguignon, Ferreira y Menéndez (2007) para estimar la desigualdad de oportunidades en Brasil y por Ferreira y Gignoux (2011) para los casos de Brasil, Colombia, Ecuador, Guatemala, Panamá y Perú. Con respecto a sus resultados, estos últimos autores señalan que deben interpretarse como un límite inferior del nivel de desigualdad de oportunidades dado que, por los problemas de estimación mencionados, se han omitido circunstancias observables, además de las no observables. Serio (2011) también utiliza esta metodología para estimar la desigualdad de oportunidades de Argentina en base a la EPH, con una selección de las variables bastante parecida a la que se realiza en este trabajo.

Este tipo de metodologías se adapta mejor a la utilización de variables continuas que el método no paramétrico descrito previamente. Sin embargo, las variables categóricas requieren un tratamiento especial. En particular, a la hora de incluir variables categóricas que pueden tomar *n* realizaciones distintas, se deben incluir en la regresión *n-1* variables binarias para representar todos los valores posibles. Esto significa que si, por ejemplo, se quiere considerar la circunstancia Raza con tres valores posibles, {Blanca; Negra; Otra}, se deben estimar dos parámetros adicionales en la regresión. De esta forma, incluir numerosas variables categóricas con varios valores posibles para cada una, puede requerir la estimación de una gran cantidad de parámetros, lo cual está limitado por el tamaño de la muestra. Es decir, al igual que con los métodos no paramétricos, puede requerirse que el investigador seleccione un subconjunto de circunstancias o limite la cantidad de valores que cada una de ellas puede tomar.

Brunori et al. (2016) indican que las estimaciones paramétricas están expuestas a los mismos sesgos que las no paramétricas. En primer lugar, a menos que se incluyan todas las circunstancias que afectan a los ingresos, las estimaciones de la desigualdad de oportunidades están sesgadas a la baja[3]. Pero intentar reducir este sesgo incrementando el número de circunstancias más allá de lo que el tamaño de la muestra puede soportar, termina introduciendo un sesgo al alza. Esto se debe a que, al incluir más circunstancias, cada tipo queda conformado por un menor número de observaciones, aumentando tanto la desigualdad entre los tipos como la varianza de las estimaciones del ingreso de cada tipo. En sucesivos trabajos, estos autores realizan dos propuestas metodológicas para superar estos inconvenientes, las cuales se presentan en la siguiente sección.

[3] Tanto Brunori et al. (2016) como Ramos y Van de Gaer (2012) sostienen que este es el caso general, aunque adminten que, en determinadas circunstancias, el sesgo por circunstancias omitidas puede ser positivo.

3.3.3 Hacia una nueva metodología de estimación

En el citado trabajo de 2016, Brunori et al. no abandonan por completo el método de estimación paramétrica explicado en la sección previa, pero introducen mecanismos de validación cruzada[4] para seleccionar la mejor especificación del modelo. Si bien esta metodología ya podría considerarse un algoritmo de ML, dado que la especificación del modelo está determinada por los datos, como se discutirá en el Capítulo 4, es una extensión del modelo lineal explicado previamente.

El método de validación cruzada consiste en dividir la muestra en *k* partes de igual tamaño, cada una de las cuales es utilizada alternativamente como conjunto de prueba o *test set*. Es decir, primero se estima el modelo utilizando las observaciones de *k-1* partes y se realiza una predicción fuera de la muestra[5] sobre el grupo de datos que no se incluyó en la estimación. Se repite el proceso *k* veces, dejando afuera de la estimación a un grupo de datos distinto cada vez, y se promedia el error de pronóstico cometido, por ejemplo, promediando el error cuadrático medio de las *k* predicciones. Este proceso se realiza para las distintas especificaciones del modelo que se quieren comparar, seleccionándose la que produce un menor error promedio. Los autores utilizan este mecanismo para 26 países europeos, probando distintos conjuntos de circunstancias e interacciones entre las mismas. Encuentran que el ranking entre los países varía dependiendo la especificación utilizada, por lo que concluyen que es extremadamente importante introducir un criterio estadístico para seleccionar un modelo entre distintas especificaciones posibles.

Brunori et al. (2018) abandonan el método de estimación paramétrico basado en regresiones lineales, adoptando los árboles de regresión condicionales que serán explicados en el Capítulo 4. Según estos autores, las ventajas de esta metodología son varias. Además de reducir aún más la influencia del investigador en la selección del modelo y las variables a utilizar, esta metodología reduce el sesgo a la baja de las circunstancias omitidas, ya que permite considerar un gran número de circunstancias y valores posibles para las mismas sin temor a un excesivo particionamiento de la muestra. En efecto, el algoritmo es el encargado de decidir, mediante criterios estadísticos, cuántas particiones son las adecuadas y en base a qué variables y valores determinarlas. Adicionalmente, se elimina la restricción de una influencia lineal y aditiva de las variables, pudiéndose captar interacciones complejas entre las mismas. Por último, esta metodología incluye mecanismos para reducir la varianza de las

[4] Traducción de *cross-validation*..
[5] Traducción de *out-of-sample*.

estimaciones, eliminando el sesgo al alza provocado por el aumento de la varianza que se origina al considerar más variables.

Comprender las particularidades de esta metodología requiere varios pasos previos, por lo que se dedicará todo el Capítulo 4 a esta tarea. En el Capítulo 6 se intentará mostrar evidencia de que efectivamente esta metodología es superadora, utilizándola luego para estimar la desigualdad de oportunidades en Argentina.

4. "Machine learning" y métodos basados en árboles

Con el aumento del poder de cómputo de los ordenadores personales y la amplia disponibilidad de datos, se popularizaron una serie de algoritmos y métodos estadísticos computacionalmente intensivos, englobados bajo el difuso nombre de 'Machine Learning' (ML). Sin embargo, no existe una única y precisa definición de ML. Lantz (2013) define a ML como el estudio y desarrollo de algoritmos computacionales para transformar datos en acción inteligente[6], mientras que en textos de mayor rigurosidad estadística como Friedman, Hastie y Tibshirani (2001) o Efron y Hastie (2016) ni siquiera se define el término. Sin aspirar a dar aquí una definición precisa sobre el término, en el contexto de este trabajo interesa establecer en qué se diferencian estos métodos de las tradicionales técnicas econométricas.

Sosa Escudero (2019), quien adopta el término 'aprendizaje automático' como traducción de ML, sostiene que estos métodos caen en la frontera entre la computación y la estadística. Afirma que, en la vieja visión de la estadística, la idea era 'estimar' un modelo propuesto por una teoría o experiencia previa: el modelo viene de afuera y los datos se usan sólo para estimarlo. Pero la profusión de datos permite construir, estimar y reestimar el modelo a medida que se lo usa. Esta es la idea detrás de 'aprender' o *learn* en lugar de 'estimar'. Por otro lado, la parte de "automático" o *machine* tiene que ver con que una parte o toda la tarea de reconstrucción del modelo puede relegarse a un procedimiento computacional.

En el mismo sentido, Athey (2018) sostiene que una característica común de los métodos ML es que la selección del modelo está basada en los datos. El analista provee la lista de variables y ejecuta un algoritmo que estima y compara numerosos -usualmente cientos o miles- de modelos alternativos, seleccionándose el mejor de acuerdo a algún

[6] "The field of study interested in the development of computer algorithms to transform data into intelligent action is known as machine learning". Lantz (2013, página 3)

criterio a maximizar. Si bien esto podría en principio generar problemas de sobreajueste[7], se utilizan distintas técnicas para reducir este problema. Esta forma de proceder contrasta con la práctica habitual de la econometría, donde el investigador elige un modelo basado en teoría y realiza una sola estimación, o unas pocas similares para testear la robustez del modelo.

Un punto importante a destacar es que los métodos de ML no agregan mucho a la economía tradicional respecto a problemas de identificación, como podría ser la identificación de un efecto causal. Es decir, no suelen ser los métodos adecuados para encontrar una buena estimación del β detrás de la relación entre X e Y. Por otro lado, suelen producir mejores predicciones en estimaciones semi-paramétricas o cuando el número de variables explicativas es grande respecto al número de observaciones (Athey, 2018). Citando a Mullainathan y Spiess (2017), ML no sólo provee nuevas herramientas, sino que resuelve un problema diferente. ML produce predicciones de Y en base a X, descubriendo patrones generalizables. En efecto, gran parte del éxito de ML se debe a su capacidad para descubrir complejas estructuras o interrelaciones que no habían sido especificadas en un modelo. Adicionalmente, al seleccionarse la forma funcional del modelo en base a los datos, permite al investigador una mayor transparencia a la hora de explicar la modelización utilizada.

Los métodos de ML se califican en 'supervisados' y 'no supervisados'. Estos últimos, que se centran en la búsqueda de patrones en las observaciones, intentando encontrar clusters o grupos de observaciones similares, no serán tratados en este trabajo. Los métodos 'supervisados' son aquellos en los que se puede distinguir claramente entre una variable objetivo (Y) y las variables explicativas (X), intentando construir un estimador $\hat{\mu}(x)$ de $\mu(x) = E[Y|X = x]$ con el que obtener buenas predicciones de Y en un set de datos independiente. En este trabajo se utilizará un tipo particular de algoritmos de ML, aquellos basados en árboles de decisión. Las siguientes secciones de este capítulo están destinadas a describir este tipo de algoritmo, desde el algoritmo base, hasta las versiones implementadas en este trabajo.

4.1 Métodos de regresión basados en árboles

Los métodos basados en árboles son muy populares en una gran variedad de disciplinas, como la investigación genética (Goldstein, Polley y Briggs, 2011; Brieuc, Waters, Drinan y Naish, 2018), la detección de spam (DeBarr y Wechsler, 2009) o incluso

[7] Traducción de *overfitting*.

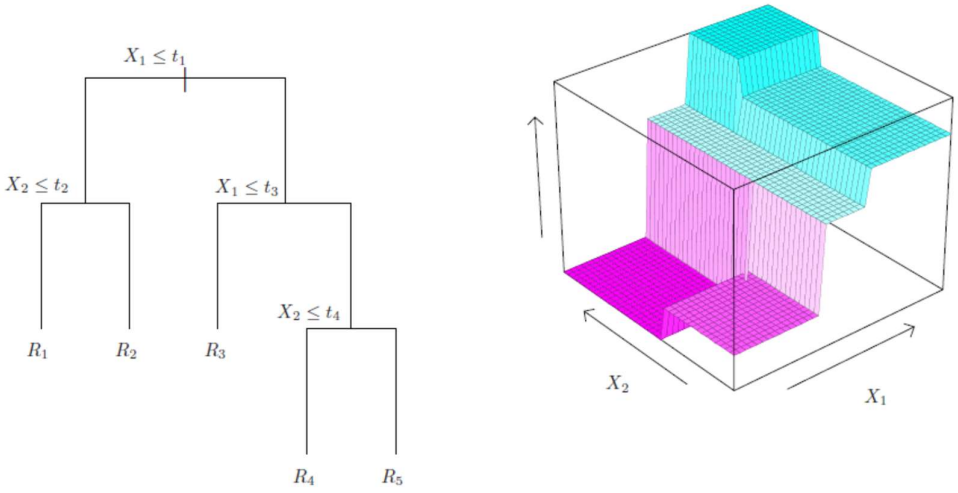
la búsqueda de planetas fuera del sistema solar (McCauliff, Jenkins, Catanzarite, Burke, Coughlin, Twicken y Cote, 2015). Se han utilizado para diversas aplicaciones financieras tales como la detección de transacciones fraudulentas (Liu, Chan, Kazmi y Fu, 2015), la predicción de defaults en préstamos (Zhou y Wang, 2012) y hasta para predecir la evolución de los mercados financieros (Khaidem, Saha y Dey, 2016; Gholamian y Davoodi, 2018). A la hora de predecir niveles de pobreza, un documento de trabajo del Banco Mundial muestra que las regresiones basadas en árboles producen resultados más precisos que otras metodologías (Sohnesen y Stender, 2017). En efecto, gran parte de la popularidad de los métodos basados en árbol se debe a que han mostrado un gran poder predictivo en ciertas situaciones donde otros tipos de modelos no se desempeñan bien o son difíciles de aplicar. Estas situaciones serían las siguientes:

- Cuando el número de predictores es muy grande respecto al tamaño de la muestra, lo cual dificulta la estimación paramétrica.
- Cuando parametrizaciones sencillas como modelos lineales o cuadráticos no son una buena aproximación al verdadero modelo generador de los datos.
- Cuando existen interacciones complejas entre las variables.
- Cuando entre las variables predictivas se incluyen tanto variables categóricas como numéricas.

En concreto, los modelos basados en árboles son modelos no paramétricos que aplican el particionamiento recursivo binario[8] sobre la muestra. Es decir, la muestra original se divide inicialmente en dos grupos en base a una variable y valor a determinar, repitiendo el proceso sucesivamente para cada submuestra hasta que se cumple cierto criterio de terminación. Como resultado, la muestra queda dividida en una cierta cantidad de grupos o nodos que tienen cierto grado de homogeneidad, lo que permite predecir un valor único de la variable dependiente para cada nodo. El siguiente esquema, tomado de Efron y Hastie (2016), ilustra el funcionamiento del algoritmo a la hora de predecir una variable Y mediante dos variables explicativas X_1 y X_2.

[8] Traducción de "binary recursive partitioning".

Gráfico 1. Esquema de los métodos de árbol



Fuente: Efron y Hastie (2016)

El diagrama de la izquierda, el "árbol", indica que para aquellas observaciones en las que se cumple tanto que $X_1 \leq t_1$ como $X_2 \leq t_2$, se predice un valor igual a R_1. De manera similar, para el resto de las observaciones se predicen valores de R_2 a R_5 dependiendo de los valores que tomen las variables X_1 y X_2. Es decir, la superficie de regresión, gráfico de la izquierda en la Figura 2, está conformada por sólo 5 valores, lo que explica su forma escalonada.

Efron y Hastie (2016) indican que un algoritmo puede tener tres usos: 1. Predicción, 2. Estimación o 3. Explicación. Los autores sostienen que los métodos basados en árboles son fáciles de interpretar, por lo que son particularmente útiles para el tercero de estos usos, el de la explicación. Sin embargo, su uso más importante es en el campo de la predicción, ya que sus errores de pronóstico suelen ser menores a los de otras metodologías, aunque con gran varianza, como se explicará más adelante. Por otro lado, las superficies de regresión discontinuas que producen los descalifica como métodos válidos para la estimación de la verdadera superficie de regresión. De todas formas, algunos métodos más sofisticados que se verán en las siguientes secciones consiguen estimar superficies más suaves.

En este trabajo la atención está puesta en los árboles de regresión, ya que lo que se intenta predecir es una variable numérica (ingreso). Pero todos los métodos explicados en este capítulo son fácilmente adaptables a casos de clasificación, es decir, cuando la variable que se quiere predecir es categórica. Para una explicación más profunda sobre árboles de regresión y de clasificación, puede recurrirse al trabajo de Strobl, Malley y Tutz (2009).

En la sección 4.2 se explicará con mayor detalle cómo es el proceso de construcción de un árbol de regresión, base de los modelos más complejos que suelen utilizarse en la mayoría de los trabajos empíricos. Las siguientes secciones estarán destinadas a explicar cómo diversas mejoras secuenciales a los árboles de regresión mejoran tanto sus propiedades estadísticas como su poder predictivo. La sección 4.3 describirá la técnica de bagging[9], mientras que en la sección 4.4 se describirán los bosques aleatorios[10], probablemente el método basado en árboles más difundido. La sección 4.5 está dedicada a los métodos que son utilizados en este trabajo: los árboles de regresión condicionales, a partir de los cuales se pueden construir bosques aleatorios condicionales. La última sección de este capítulo explica cómo se pueden utilizar estos modelos para estimar la desigualdad de oportunidades.

4.2 Árboles de Regresión

Supongamos que se dispone de P variables explicativas para estimar el valor de una variable numérica Y, contando para ello con N observaciones (X_i, Y_i) con $i = 1,2, \dots, N$ y $X_i = (X_{i1}, X_{i2}, \dots, X_{iP})$. Si bien existen distintos algoritmos para crear un árbol de regresión, uno de los más utilizados es el desarrollado por Breiman et al. (1984), conocido como el algoritmo CART[11], que se describe a continuación.

Partiendo de la muestra completa, se considera la variable explicativa X_j y el valor s para definir las siguientes particiones:

$$R_1(j,s) = \{X | X_j \leq s\} \; y \; R_2(j,s) = \{X | X_j > s\}$$

Para cada una de estas particiones se predice el valor de Y simplemente tomando el promedio de todos los Y_i de cada partición.

$$\hat{y}_k = \sum_{i \in R_k} {}^{y_i}/_{N_k}$$

El algoritmo prueba con diversos valores de j y s, eligiendo aquellos que minimizan las sumas del error de predicción cuadrático de cada partición:

$$s_1^2 + s_2^2$$

Siendo $s_k^2 = \sum_{i \epsilon R_k} (y_i - \hat{y}_k)^2$. Puede demostrarse que minimizar esta suma equivale a maximizar la diferencia entre las medias de cada grupo (Efron y Hastie, 2016), por lo que el algoritmo elige *j* y *s* de tal modo que los dos grupos de datos son lo más diferente posible.

[9] "Bagging" suele traducirse como "agregación" o "empaquetado". Para evitar confusiones, se decidió mantener la terminología en inglés.
[10] Traducción de "random forest".
[11] Sigla de "Classification and Regression Trees": árboles de regresión y clasificación.

Por ejemplo, un particionamiento ideal sería cuando todos los elementos de un grupo son 0 y todos los elementos del otro grupo son 1, por lo que la varianza de cada partición es 0.

Una vez que se dividió la muestra original en dos submuestras o ramas del árbol, se procede de manera similar para cada rama, es decir, al interior de cada rama se vuelve a seleccionar una variable *j* y un valor *s* para volver a dividir esa porción de la muestra en dos submuestras más pequeñas, dando lugar a nuevas ramas del árbol. ¿Hasta dónde se deja que crezca el árbol? Un árbol muy grande produciría un sobreajueste sobre la muestra de entrenamiento, generando malos pronósticos fuera de la muestra. Por otro lado, un árbol demasiado corto podría no capturar parte de la estructura presente en los datos. El criterio más común consiste en dejar crecer un árbol muy largo, por ejemplo, hasta que cada grupo tenga menos de x elementos, para luego podarlo[12] mediante técnicas de validación cruzada y análisis de costo-complejidad. Dado que en este trabajo se utilizarán árboles de regresión condicionales, los cuales utilizan criterios de significancia estadística para dejar de dividir la muestra, no se profundizará en los detalles del proceso de poda o *prunning* (Ver Efron y Hastie (2016) o Friedman et al (2001) para más detalles).

Dos grandes ventajas de los árboles de regresión son su interpretabilidad (Friedman et al, 2001) y la capacidad de captar interacciones complejas entre las variables (Strobl et al, 2009). Pero presentan al menos dos grandes problemas. Por un lado, como ya fue mencionado la superficie de predicción carece de suavidad. Por otro lado, las predicciones de los árboles son muy inestables. Un pequeño cambio en la muestra puede generar particiones muy diferentes, por lo que son estimadores de varianza elevada. Este es el problema más relevante cuando el objetivo es conseguir buenas predicciones. Afortunadamente, la varianza de las estimaciones puede reducirse construyendo un ensamble de una gran cantidad de árboles. Esta técnica, conocida como 'bagging', se describe en la siguiente sección.

4.3 Bagging

Diversos estudios empíricos o basados en simulaciones han mostrado que los árboles de clasificación producen, en promedio, la predicción correcta. Es decir, son predictores insesgados (Bauer y Kohavi, 1999; Breiman, 1996; Breiman, 1998; Dietterich, 2000; Bühlmann y Yu, 2002). Si bien existe menos evidencia para el caso de los árboles de regresión, la literatura considera que son aproximadamente insesgados y que, en todo caso,

[12] En inglés, al proceso de poda se lo suele llamar *"prunning"*.

el sesgo de un ensamble de árboles es el mismo que el de un árbol individual (Friedman et al, 2001). Es decir, el ensamble de muchos árboles reduce la varianza o inestabilidad de las estimaciones, haciendo más confiables las predicciones, sin modificar el sesgo. Ahora bien, dado que sólo se cuenta con un conjunto de datos ¿de dónde salen todos estos árboles diferentes? El primer paso consiste en construir distintos conjunto de datos mediante la técnica de *bootstrapping*.

Se entiende por *bootstrap* al proceso de construir una muestra de igual tamaño a la original, digamos N, mediante extracciones con reposición de una muestra de N observaciones de la muestra original. Es decir, de la muestra original conformada por N pares de valores (x_i, y_i), se extraen N valores con reposición, dando lugar a una muestra de igual tamaño a la original pero, casi con seguridad, ligeramente diferente, dado que una misma observación puede ser elegida más de una vez.

Para realizar el proceso de ensamble o *bagging*, se obtienen B muestras mediante *bootstrap*[13] y para cada una de ellas se corre un árbol de regresión. Dada la inestabilidad de los árboles y las diferencias entre las muestras, los árboles estimados pueden diferir en muchos aspectos: las variables elegidas, los valores de corte y en la cantidad de nodos terminales identificados. Lo usual es dejar crecer árboles grandes, sin podarlos, dando lugar a árboles mucho más diversos, con distintas combinaciones de las variables predictivas. Como cada uno de estos árboles produce su propia predicción de y para cada observación de x, a la que denominaremos $\hat{y}^b(x)$, el estimador *bagged* se construye mediante el promedio de todas estas estimaciones:

$$\hat{y}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{y}^b(x)$$

Este método no sólo reduce significativamente la varianza de las estimaciones, sino que también suaviza la superficie de predicción (Friedman et al, 2001). Esto se debe a que cada árbol conforma los grupos de manera diferente, por lo que ya no existen k grupos y k predicciones para la variable dependiente, pudiendo existir hasta una predicción por observación.

La consistencia de estos estimadores ha sido analizada por Breiman (2001). Este autor prueba que, cuando el número de árboles ensamblados se incrementa, el error de predicción converge a un error generalizado, cuyo límite superior depende positivamente de

[13] Lo más usual es utilizar *bootstrap*, aunque también podrían construirse submuestras de menor tamaño a la original, extraídas sin reposición.

la correlación entre los árboles. Es decir, si se ensamblan árboles parecidos, el algoritmo convergerá hacia un error generalizado alto. Si se ensamblan árboles muy diferentes, el algoritmo convergerá hacia un error menor. Esto tiene dos importantes consecuencias. En primer lugar, el ensamble de árboles no produce sobreajuste, dado que por más que se sigan adicionando árboles no se puede reducir el error más allá de cierto límite. En segundo lugar, si se consigue reducir la correlación entre los árboles, se conseguiría reducir el error de predicción. Una forma de hacer esto es agregar una fuente más de aleatoriedad en la selección de las variables de cada árbol, que es lo que realizan los bosques aleatorios analizados en la siguiente sección.

4.4 Bosques Aleatorios[14]

El método de regresión basado en árboles más difundido es sin duda el de bosques aleatorios, dado el buen desempeño predictivo que muestra en diversos contextos. Introducidos por Breiman (2001), podrían entenderse como un proceso de *bagging*, tal como se describió en la sección anterior, al que se le agrega una fuente adicional de aleatoriedad destinada a romper la correlación entre los distintos árboles ensamblados. En particular, si contamos con P variables explicativas, el truco consiste en no considerarlas a todas juntas todo el tiempo. Más específicamente, en cada nodo de cada uno de los árboles ensamblados, se selecciona al azar un grupo $m < P$ de variables para considerar a la hora de dividir la muestra[15]. Esto da lugar a árboles aún más diversos que los considerados bajo *bagging*.

Por ejemplo, si se dispone de cinco variables explicativas $X_1, \dots, X_5$ y se decide trabajar con un valor de $m = 2$, el azar puede determinar que para el primer nodo del árbol sólo se consideren las variables X_1 y X_2, mientras que para el segundo nodo sólo se consideren las variables X_3 y X_5 y así sucesivamente. Y para cada árbol del ensamble, el azar determinará distintos grupos de variables para cada nodo.

La primera consecuencia de este proceso es que se reduce considerablemente la correlación entre los árboles, ya que los mismos están obligados a considerar distintas variables en cada división de la muestra. Como se mencionó en la sección anterior, esto reduce el límite superior del error generalizado que produce el ensamble de árboles a la hora de predecir. Es decir, permite la convergencia hacia un menor error de predicción.

[14] Traducción de *random forest*.
[15] También podría permitirse $m = p$, aunque en este caso el bosque aleatorio se reduce al método de *bagging*.

Otra consecuencia es que permite una mejor evaluación sobre la contribución de cada variable a la hora de predecir la variable explicativa. Imaginemos dos variables explicativas X_1 y X_2, ambas con influencia sobre Y pero que correlacionan entre sí. Al estimar un árbol, supongamos que el algoritmo encuentra que la mejor forma de dividir la muestra inicial es en base a la variable X_1. De ahí en más, puede suceder que X_2 no vuelva a ser considerada ya que parte de su influencia ya fue captada por X_1, quedando relegada por otras variables en las sucesivas subdivisiones de la muestra. En un bosque aleatorio, por el contrario, existirán árboles donde el azar determinará que se considere la variable X_2 y otras más, pero no X_1, permitiendo a X_2 mostrar su poder predictivo y dando lugar a interacciones más complejas que las permitidas por otras metodologías.

La cantidad m de variables a considerar en cada nodo es un parámetro a estimar, aunque valores típicos para bosques de regresión van desde 1 a $P/3$, donde P es la cantidad de variables. Más allá de que el valor elegido de m puede resultar del testeo de varios valores alternativos[16], existen algunos aspectos a tener en cuenta. Friedman et al (2001) indica que, cuando el número de variables es grande pero la fracción de variables relevantes es pequeña, la performance de los bosques aleatorios no es muy buena para valores muy bajos de m. Esto se debe a que se reduce considerablemente la probabilidad de que en un nodo determinado sea elegida una variable relevante. Por el contrario, cuando el número de variables relevantes se incrementa, la performance de los bosques aleatorios mejora considerablemente. Efron y Hastie (2016) indica que elegir $m = 1$ reduce al máximo la correlación entre las árboles pero puede introducir algún sesgo en las estimaciones.

Más allá de que, a la hora de predecir, los bosques aleatorios son sensiblemente superiores a los árboles de regresión, no son tan fáciles de interpretar. En efecto, no es posible dibujar un 'árbol promedio', dado que en cada uno de los árboles del ensamble pueden aparecer distintas variables en distintas posiciones. Es por eso que lo usual es combinar ambas metodologías. Los árboles de regresión se utilizan para explicar e interpretar los hallazgos del modelo, mientras que las predicciones y la importancia de cada variable se estima mediante los bosques aleatorios. Respecto a la importancia de las variables, es quizá el principal resultado que nos da un bosque aleatorio en términos de interpretabilidad.

Existen distintas formas de medir la importancia de las variables. De acuerdo a Strobl et al (2009), la más sofisticada de estas medidas es la basada en permutaciones y predicciones

[16] Proceso conocido como *tunning* de los parámetros.

out-of-bag[17] (OOB). Recordemos que cada árbol del ensamble se corre en una muestra generada mediante *bootstrapping*, lo que significa que algunas observaciones no son consideradas para construir los árboles. En particular, puede probarse que el valor esperado del porcentaje de valores únicos extraídos mediante *bootstrapping* de una muestra de *N* elementos es:

$$1 - \left(1 - \frac{1}{N}\right)^N$$

Para valores de *N* suficientemente grandes, este valor converge a:

$$1 - \frac{1}{e} \approx 0.632120559\ldots$$

Es decir, para valores de *N* suficientemente grandes se espera que alrededor del 63% de la muestra sea considerada en cada árbol, dejando un 37% (Efron y Hastie, 2016) para realizar predicciones OOB. Las predicciones OOB consisten en realizar la predicción correspondiente a la observación $z_i = (x_i, y_i)$, promediando sólo las predicciones de los árboles en los que z_i no fue utilizado para la estimación. De esta forma, se obtiene una estimación más realista del error que se espera obtener en una nueva muestra, ya que sólo se están considerando predicciones fuera de la muestra.

Para estimar la importancia de las variables en base a permutaciones, se procede de la siguiente manera. Luego de estimar un árbol específico, se estima el error de predicción en las muestras OOB. A continuación, los valores de la variable cuya importancia se desea estimar, X_j, son permutados al azar, y se repite la estimación del error de predicción en las muestras OOB. La diferencia entre ambos errores de predicción se promedia a lo largo de todos los árboles, dando lugar a una medida de la importancia de la variable X_j en el bosque aleatorio. La idea detrás de este mecanismo es que, si existe alguna relación entre la variable X_j y la variable dependiente, esta relación se rompe al permutar los valores de la variable X_j. De esta forma, se está midiendo cuánto de la contribución de la variable X_j a mejorar el error de predicción puede atribuirse a algo más que azar.

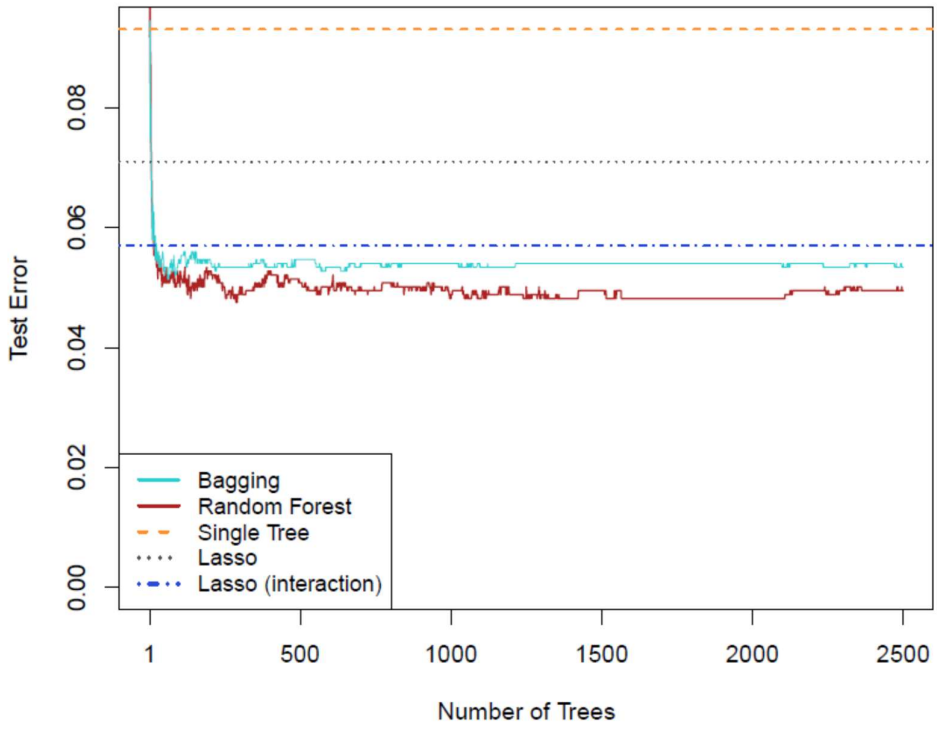
Más allá de los aspectos técnicos de la medida de importancia de las variables, debe quedar en claro que este indicador no es un porcentaje y, por lo tanto, pueden surgir valores negativos. En efecto, puede suceder que una variable no esté relacionada con la variable dependiente o la relación sea muy débil. En este caso, puede ocurrir que, por simple azar, la permutación de los valores de la variable en cuestión mejore ligeramente la predicción

[17] Se decidió utilizar el término en inglés, el cual hace referencia a las observaciones que quedan fuera en el proceso de *bagging*.

respecto a la variable original, dando como resultado un valor negativo para la importancia de la variable. Como consecuencia de esto, variables cuya medida de importancia resulta en una cantidad positiva pero del mismo orden de magnitud que los valores negativos de otras variables, no deberían considerarse importantes para el modelo, ya que no puede descartarse que la mejora que producen en el error de estimación sea por efectos del azar.

Un aspecto final a destacar sobre los bosques aleatorios es que, tal como su nombre lo indica, son aleatorios. Es decir, si se ejecuta dos veces el algoritmo, seguramente se obtendrán resultados ligeramente diferentes. Las dos fuentes de aleatoriedad son: las muestras conformadas mediante *bootstrap*, y la selección aleatoria de las variables que se pueden considerar en cada nodo. Una tercera fuente de aleatoriedad surge a la hora de estimar la importancia de las variables, con la permutación aleatoria de las variables explicativas. De todas formas, si el número de árboles ensamblados es lo suficientemente grande, la diferencia entre distintas ejecuciones del algoritmo debería ser completamente despreciable, dada la convergencia hacia cierto valor o error generalizado demostrada por Breiman (2001). Esta convergencia hacia un nivel de error generalizado puede ilustrarse con la siguiente figura extraída de la comparación de distintos modelos para predecir si un correo electrónico es spam o no.

Gráfico 2. Estabilización del error de pronóstico para un número grande de árboles



Fuente: Efron y Hastie, 2016

Puede observarse que el error de pronóstico de los métodos de *bagging* y bosques aleatorios cae abruptamente en los primeros cien árboles, mejorando sensiblemente los resultados de un único árbol. Pero de ahí en más el error se estabiliza, por lo que no se gana nada con ensamblar 2500 árboles en lugar de 200. Estos valores pueden variar para cada aplicación en particular, pero lo que se quiere resaltar es que, usualmente no se gana nada con ensamblar un millón de árboles respecto al ensamble de un número mucho más reducido de árboles.

Si bien los bosques aleatorios son el método más relevante de los métodos de ML basados en árboles, el contexto de este trabajo amerita la utilización de una metodología ligeramente diferente, que incorpora nociones de significancia estadística en cada nodo del árbol. La siguiente sección explica los detalles de esta metodología.

4.5 Árboles y bosques de regresión condicionales

Un aspecto de los árboles de regresión sobre el que no existe consenso absoluto es respecto a cuándo el algoritmo debe detenerse y no seguir dividiendo la muestra. Como se explicó en las secciones anteriores, el método usual es estimar árboles grandes y luego podarlos mediante pruebas de validación cruzada, con el objeto de reducir el sobreajuste. Hothorn et al (2006) recogen una crítica de Mingers, quien afirmó que el algoritmo *"no tiene noción de significancia estadística y, por lo tanto, no puede distinguir entre una mejora significativa de una no significativa"* en la estimación. Adicionalmente, los mismos autores mencionan que la literatura ha identificado que, a la hora de elegir la variable en base a la cual dividir la muestra, los métodos descritos previamente tienen un sesgo a favor de las variables que tienen muchas observaciones faltantes. Si bien este último aspecto no es relevante para este trabajo, dado que sólo se trabaja con observaciones en los que hay datos para todas las variables, incorporar elementos de significancia estadística sí lo es para poder rechazar o no la hipótesis nula de existencia de igualdad de oportunidades. A continuación, se presenta la metodología propuesta por estos autores, la cual es una modificación del algoritmo CART explicado en la sección 4.2.

La novedad que incorporan los árboles de regresión condicionales es que utilizan un test de hipótesis a la hora de decidir si el árbol debe seguir creciendo o no. En concreto, en cada nodo se adopta la hipótesis nula de que la distribución de la variable dependiente Y es independiente de las P variables explicativas. En particular, hay una hipótesis nula parcial para cada variable explicativa, dando lugar a la hipótesis nula global, la cual se cumple cuando se cumplen todas las hipótesis parciales:

$$p\ Hipótesis\ Nulas\ parciales: H_0^j: D\left(Y \middle| X_j\right) = D(Y)$$

$$Hipótesis\ Nula\ global: H_0 = \cap_{j=1}^{p} H_0^j$$

Cuando no es posible rechazar la hipótesis nula global con un nivel de significatividad α predeterminado, el árbol deja de dividirse. Si se puede rechazar la hipótesis nula global, entonces se testean todas las hipótesis parciales, escogiéndose la variable X_j cuyo test produce el menor *valor-p* en el test de hipótesis individual.

El proceso por el cual se ensamblan muchos árboles para dar origen a los bosques aleatorios condicionales es similar al descrito en las secciones previas, siendo válido todo lo explicado respecto a *bagging* y aleatoriedad en la selección de las variables. También es similar la forma en que se mide la importancia de las variables en base a las permutaciones y predicciones OOB, por lo que no se repetirán aquí las explicaciones. En la siguiente sección se explica cómo puede utilizarse esta metodología para estimar la desigualdad de oportunidades, tal como se hará en este trabajo.

4.6 Utilización de árboles condicionales para estimar la desigualdad de oportunidades

En un reciente documento de trabajo del Banco Mundial (Brunori et al, 2018), se realiza una estimación de la desigualdad de oportunidades para países europeos utilizando árboles y bosques condicionales, tal como los descritos en la sección previa. Como se explicó en el Capítulo 3, los autores justifican la adopción de métodos de ML por dos grandes razones:

1. Las estimaciones de la desigualdad de oportunidades son muy dependientes del tipo de modelización empleada, siendo necesario reducir lo máximo posible las decisiones que toma el investigador en base a ideas previas. Es decir, se necesita un modelo basado en datos.
2. Los métodos de ML producen una mejor estimación de la distribución de ingresos contrafactual (ver Capítulo 3).

De acuerdo a los citados autores, los árboles de regresión condicionales y sus correspondientes bosques ofrecen una estructura particularmente relevante en el contexto de igualdad de oportunidades. Como se explicó en la sección 4.5, el algoritmo procede a dividir la muestra sólo si puede rechazar la hipótesis nula global de que la distribución de la variable dependiente es independiente de las variables explicativas. En el contexto de este trabajo, lo que se testea es si la distribución de los ingresos individuales es independiente de las

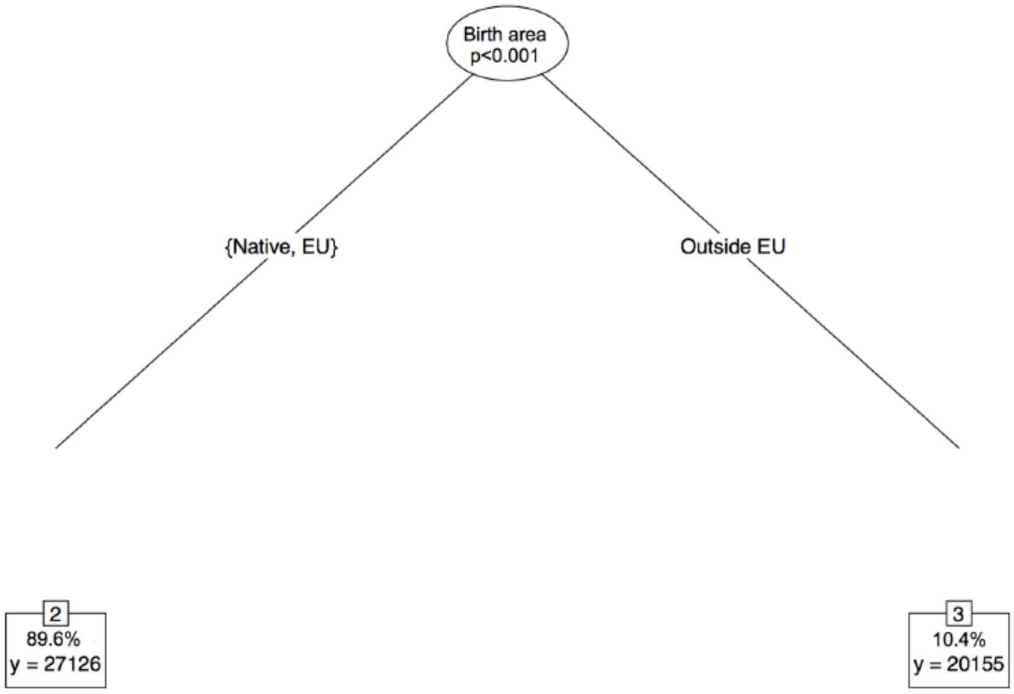
circunstancias consideradas. Es decir, adoptando la definición *ex – ante* de igualdad de oportunidades, cada test de hipótesis es esencialmente un test acerca de si existe la igualdad de oportunidades en cada submuestra particular. Si el algoritmo decide no proceder a dividir la muestra, es porque no puede rechazar la hipótesis nula de igualdad de oportunidades. Por el contrario, cada vez que el algoritmo produce una ramificación en el árbol, es porque hay evidencia significativa de que cada grupo de individuos goza de distintas oportunidades a la hora de obtener un ingreso.

El citado estudio realiza una ilustración empírica utilizando la ronda 2011 de la *European Union Statistics on Income and Living Conditions Survey*, la cual provee datos de corte transversal sobre ingreso, pobreza y condiciones de vida de 31 países europeos. Lo que hace particularmente interesante a esa ronda de la encuesta es que cuenta variables intergeneracionales, algo que por desgracia no está presente en la EPH. Las circunstancias que consideran para estimar la distribución de ingresos contrafactual son las siguientes: sexo, lugar de nacimiento, presencia de padres en el hogar, cantidad de adultos en el hogar, cantidad de adultos que trabajan en el hogar, cantidad de niños en el hogar, país de nacimiento del padre/madre, nivel educativo del padre/madre, status ocupacional del padre/madre, principal ocupacional del padre/madre y régimen de tenencia de la vivienda.

Los autores centran su análisis en dos casos en cierta medida extremos: Suecia y Alemania. A continuación, se presentan estos resultados, dado que pueden servir para interpretar los resultados para Argentina que se presentan en el capítulo siguiente.

La interpretación del árbol de oportunidades para Suecia, que se presenta en el Gráfico 3, es la siguiente. Con un valor-p de menos de 0,001, se rechaza la hipótesis nula de existencia de igualdad de oportunidades. Al comparar el valor-p de los tests de hipótesis individuales, el algoritmo selecciona a la nacionalidad de la persona como la variable de corte, procediendo a la división de la muestra en base a la misma. En particular, evalúa que el valor de corte adecuado es el de haber nacido o no en la Unión Europea: quienes nacieron en la Unión Europea obtienen un nivel de ingresos significativamente mayor a los que nacieron fuera de la Unión Europea. Una vez que se separa la población en estos dos grupos, se repite el proceso en cada rama del árbol. Pero a partir de ahí, el algoritmo ya no consigue volver a rechazar la hipótesis nula de igualdad de oportunidades. Es decir, al interior del 89,6% de la población de Suecia que nació en la Unión Europea, no hay evidencia de que exista desigualdad de oportunidades. Lo mismo puede decirse para el 10,4% de la población que nació fuera de la Unión Europea.

Gráfico 3. Árbol de Oportunidades para Suecia



Fuente: Brunori et al, 2018

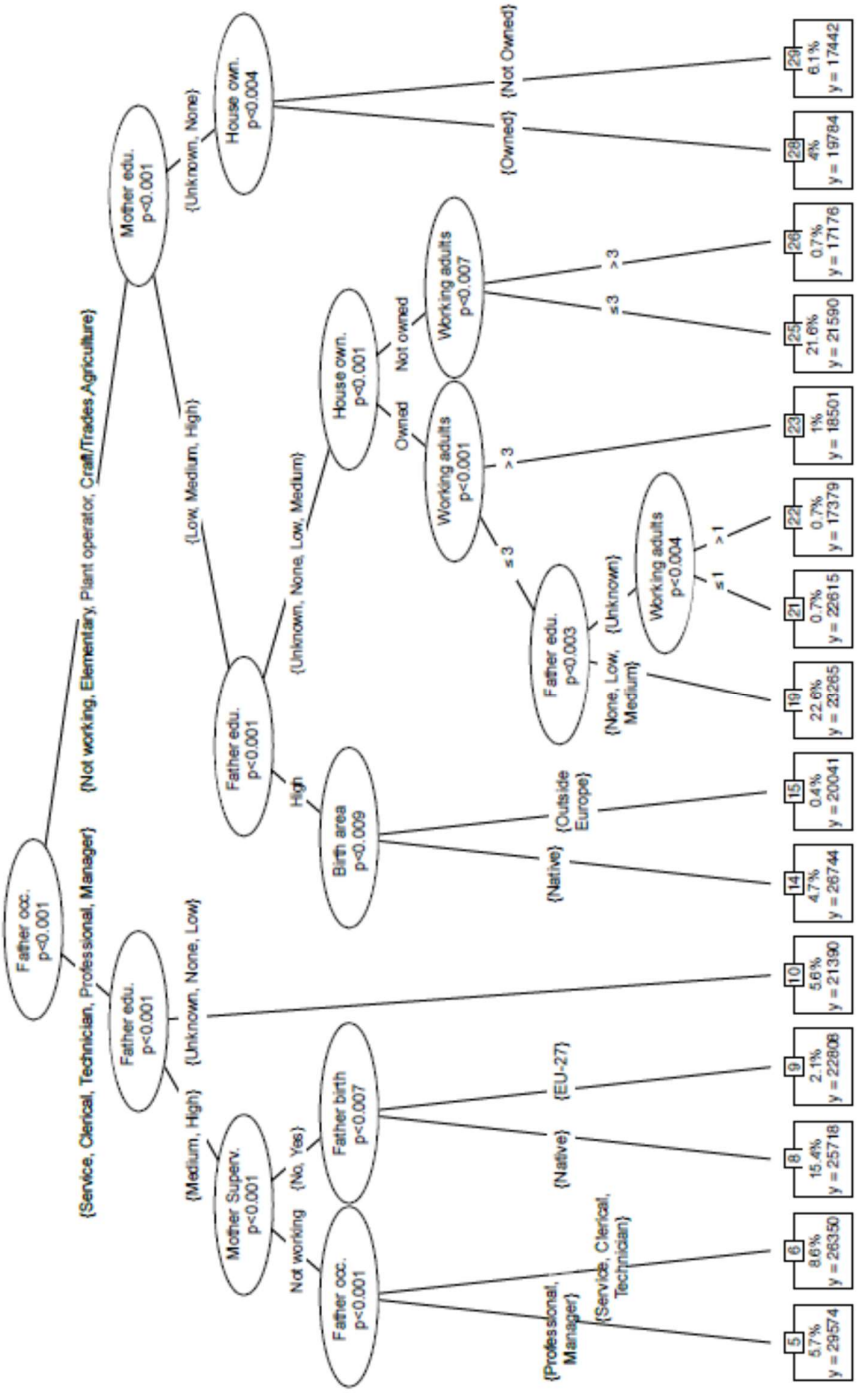
El caso de Alemania es mucho más complejo y se invita al lector a que lo analice con detalle (ver Gráfico 4). En concreto, se rechaza la hipótesis nula de existencia de igualdad de oportunidades varias veces, dando como resultado la conformación de 15 grupos o tipos de diferentes oportunidades en base a la ocupación, educación y lugar de nacimiento de los padres, entre otras variables.

El diferente grado de complejidad de los árboles de oportunidades de estos países también se observa cuando se analiza la importancia de cada variable. Para esto, ya se están utilizando los resultados de las estimaciones mediante bosques aleatorios, las cuales, recordemos, no permiten una visualización tan vistosa como la de los árboles. En el Gráfico 5 puede observarse que, mientras para el caso de Suecia sólo el lugar de nacimiento parece ser relevante, en Alemania son varias las variables que importan a la hora de generar diferentes oportunidades.

A la hora de cuantificar el nivel de desigualdad de oportunidades de estos países y de los demás que componen la muestra, los autores realizan lo siguiente. En primer lugar, realizan una predicción del ingreso de cada individuo en base a las circunstancias consideradas: la distribución de ingresos contrafactual. Las diferencias de ingreso predichas por esta estimación son la generadas por la existencia de desigualdad de oportunidades,

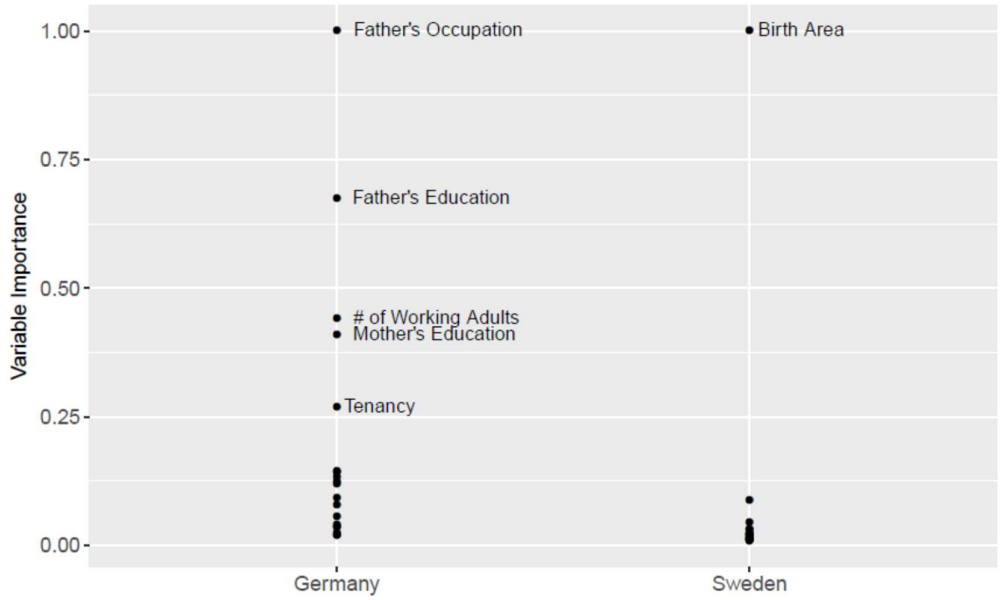
mientras que las diferencias de ingresos observadas en la población pero no explicadas por las circunstancias, podrían atribuirse a distintas decisiones o esfuerzos de los individuos. De esta manera, para estimar la desigualdad de oportunidades, los autores calculan un coeficiente de Gini para la distribución de los ingresos contrafactual.

Gráfico 4. Árbol de Oportunidades para Alemania



Fuente: Brunori et al, 2018

Gráfico 5. Importancia de las variables para Alemania y Suecia



Fuente: Brunori et al, 2018

Los resultados, que se presentan a continuación, difieren ligeramente dependiendo de si se estimaron con un árbol o con un bosque. Recordar que, al tratarse de un coeficiente de Gini, un valor de 0 corresponde a la ausencia de desigualdad, mientras que un valor más cercano a 1 indica una mayor desigualdad.

Tabla 2. Desigualdad de Oportunidades para Suecia y Alemania

	Árbol Condicional	Bosque Condicional
Suecia	0,0247	0,0313
Alemania	0,0697	0,0793

Fuente: Brunori et al, 2018

En el Capítulo 6 se intentará realizar un análisis similar para Argentina en base a la EPH. La ausencia de variables intergeneracionales en la encuesta impide un análisis tan rico como el anterior, aunque los resultados no dejan de ser relevantes. Previamente, es necesario hacer algunas consideraciones respecto a la selección de las variables, lo cual se hará en el capítulo siguiente.

5. Selección de las variables a utilizar

Todo el trabajo empírico se realizará utilizando como fuente la Encuesta Permanente de Hogares (EPH) elaborada por el INDEC, combinando variables de la base de hogares y de personas. La EPH es un programa nacional de recolección de datos con el objetivo de conocer las características socioeconómicas de la población. En base a esta encuesta se calculan las tasas oficiales de empleo, desocupación, subocupación y pobreza. Desde el año 2003, la encuesta se realiza de manera trimestral, cubriendo 32 aglomerados urbanos y alrededor de 18 mil hogares por trimestre. Los hogares son entrevistados mediante un esquema de rotación 2-2-2 (INDEC, 2003), que consiste en lo siguiente:

- Las viviendas de un área ingresan a la muestra para ser encuestadas en dos trimestres consecutivos.
- Se retiran por dos trimestres consecutivos.
- Vuelven a la muestra para ser encuestadas en dos trimestres consecutivos.

De esta forma, una vivienda que es encuestada por primera vez en el trimestre 1, vuelve a ser encuestada en el trimestre 2, se retira momentáneamente de la muestra para volver a ser encuestada en el trimestre 1 y 2 del año siguiente. Es decir, un hogar puede ser seguido a lo largo de un año y medio.

Si bien están disponibles las bases de microdatos desde el año 2003, en el año 2016 se introdujeron algunos cambios metodológicos en la encuesta (INDEC, 2016), por lo que en este trabajo se decidió utilizar sólo la información disponible desde el segundo trimestre de 2016. A la fecha en que se escribe este trabajo, la última base disponible es la del tercer trimestre de 2018, por lo que se cuenta con información para diez trimestres consecutivos.

Los microdatos de la encuesta para cada trimestre están disponibles en dos bases separadas, una a nivel hogar y otra a nivel individuo. Ambas pueden combinarse mediante un código para distinguir vivienda, otro código para distinguir hogares y, en el caso de la base de individuos, un código correspondiente a cada componente del hogar (INDEC, 2019). Además de estas variables, las correspondientes al trimestre y año de realización de la encuesta, para este trabajo se han considerado las variables que se especifican en la Tabla 3.

Como se verá en el siguiente capítulo, se realizaron estimaciones tanto a nivel país como para cada una de las seis regiones en las que la EPH subdivide al país (Gran Buenos Aires, Noroeste, Noreste, Cuyo, Pampeana y Patagonia). En todos los casos, se filtró la base de datos de manera de considerar sólo las personas de entre 30 y 59 años que tienen un ingreso mayor a cero, dado que el interés está puesto en las personas que obtienen ingresos.

Tabla 3. Variables consideradas en este trabajo.

Variables presentes en ambas bases		
Variable	**Descripción**	**Codificación**
REGION	Código de Región	01 = Gran Buenos Aires, 40 = NOA, 41 = NEA, 42 = Cuyo, 43 = Pampeana, 44 = Patagonia
MAS_500	Aglomerados según tamaño	S = Conjunto de Aglomerados de 500.000 y más hab.
Base a nivel Hogar		
II7 y II7_Esp	Régimen de tenencia	01 = Propietario de la vivienda y el terreno, 02 = Propietario de la vivienda solamente, 03 = Inquilino/arrendatario de la vivienda, 04 = Ocupante por pago de impuestos/expensas, 05 = Ocupante en relación de dependencia, 06 = Ocupante gratuito (con permiso), 07 = Ocupante de hecho (sin permiso), 08 = Está en sucesión?, 09 = Otra situación (especificar)
IX_Tot	Cantidad de miembros del Hogar	Variable numérica
IX_Men10	Cantidad de miembros del Hogar menores de 10 años	Variable numérica
Base a nivel Personas		
COMPONENTE	Número de componente: Nº de orden que se asigna a las personas que conforman cada hogar de la vivienda.	Casos especiales: 51 = Servicio doméstico en hogares. 71 = Pensionistas en hogares.
CH03	Relación de Parentesco	01 = Jefe/a, 02 = Cónyuge/Pareja, 03 = Hijo/Hijastro/a, 04 = Yerno/Nuera, 05 = Nieto/a, 06 = Madre/Padre, 07 = Suegro/a, 08 = Hermano/a, 09 = Otros Familiares, 10 = No Familiares
CH04	Sexo	1= varón, 2 = mujer
CH06	¿Cuántos años cumplidos tiene?	Variable numérica
CH15	¿Dónde nació?	1. En esta localidad, 2. En otra localidad de esta provincia, 3. En otra provincia (especificar), 4. En un país limítrofe (especificar Brasil, Bolivia, Chile, Paraguay, Uruguay), 5. En otro país (especificar), 9. N/S.N/R.
CH16	¿Dónde vivía hace 5 años?	1. En esta localidad, 2. En otra localidad de esta provincia, 3. En otra provincia (especificar), 4. En un país limítrofe (especificar Brasil, Bolivia, Chile, Paraguay, Uruguay), 5. En otro país (especificar), 6. No había nacido, 9. N/S.N/R.
NIVEL_ED	NIVEL EDUCATIVO	1 = Primaria Incompleta (incluye educación especial), 2 = Primaria Completa, 3 = Secundaria Incompleta, 4 = Secundaria Completa, 5 = Superior Universitaria Incompleta, 6 = Superior Universitaria Completa, 0 = Sin instrucción, -1 = Ns./ Nr. [18]
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
p47T	MONTO DE INGRESO TOTAL INDIVIDUAL	Variable numérica

Para estimar la desigualdad de oportunidades, no existe una única manera de seleccionar la variable objetivo ni las circunstancias ajenas al control individual. En el caso de las circunstancias, la selección de las mismas estuvo básicamente basada en la disponibilidad de los datos. Pero la selección de la variable objetivo no era trivial, razón por la cual se dedica la siguiente sección a justificar la decisión tomada.

[18] Para la variable "Nivel educativo", se modificó la codificación original para que la variable tenga un orden incremental. En la codificación de la EPH, la respuesta "Sin instrucción" se codifica con un 7 y la falta de respuesta con un 9.

5.1. La elección de la variable dependiente.

En este trabajo, la variable que se intenta estimar es el monto de ingreso total individual (variable p47T en la EPH). Esta elección es diferente a la de la literatura más cercana, por lo que necesita justificación. En particular, considero relevante justificar por qué la elección es diferente a la de los dos trabajos siguientes:

- Serio (2011), quien también estima la desigualdad de oportunidades en base a la EPH, considera como variable objetivo el ingreso laboral por hora.
- Brunori et al. (2018), quienes emplean bosques aleatorios condicionales para estimar la desigualdad de oportunidades en Europa, utilizan el ingreso del hogar por adulto equivalente.

A continuación, se presentan una serie de observaciones tendientes a justificar la decisión adoptada en este trabajo. En primer lugar, considero que el nivel de vida de una persona está asociado con el nivel total de ingresos que obtiene, independientemente de si es un ingreso laboral o de otra fuente, por lo que al estimar la desigualdad de oportunidades se deberían considerar todos los ingresos. En este sentido, en la EPH se consideran bajo ingresos no laborales, ingresos por alquiler de propiedades e ingresos por intereses de inversiones, entre otros, los cuales probablemente sean tanto o más significativos para explicar la desigualdad de oportunidades que los ingresos laborales. Por otro lado, dentro de ingresos no laborales también se incluyen seguros de desempleo y otro tipo de ayuda social que muchas veces van destinados a personas que no tienen ingresos laborales. Los argumentos para incluir este tipo de ingresos son básicamente dos. Por un lado, estos ingresos no laborales suelen ser menores que los ingresos laborales, al menos en el mercado formal. De esta forma, considero que incluir sólo los ingresos laborales dejaría afuera a personas que están fuera del mercado laboral de manera involuntaria, subestimando el grado de desigualdad de oportunidades. Por otro lado, de no considerar este tipo de ingresos se estaría ignorando el efecto igualador de oportunidades que tienen determinadas políticas. En este sentido, puede decirse que la desigualdad de oportunidades que se estima en este trabajo ya considera la intervención de ciertas políticas públicas destinadas a reducir las desigualdades de ingresos.

En segundo lugar, respecto a considerar el ingreso mensual o el ingreso por hora, es discutible hasta qué punto una persona puede elegir cuántas horas trabajar. Por ejemplo, una persona puede tener un salario por hora relativamente elevado, pero aun así estar

subempleado de manera involuntaria, obteniendo bajos ingresos mensuales. Por estos motivos, considero que el ingreso total que percibe un individuo en un mes es más representativo que el ingreso laboral por hora.

En tercer lugar, utilizar el ingreso del hogar por adulto equivalente tal como hacen Brunori et al., requiere descartar dos circunstancias que en principio parecen relevantes: género y tamaño del hogar. El efecto de la variable género se diluye porque tanto el jefe de hogar como su cónyuge tienen el mismo ingreso del hogar por adulto equivalente. Por otro lado, aún utilizando las tablas de equivalencia por adulto equivalente oficiales (INDEC, 2018) para construir esta variable, se observa que el ingreso del hogar por adulto equivalente correlaciona fuertemente y de manera negativa con el tamaño del hogar. En efecto, si se intenta correr un árbol de regresión con esta variable, el algoritmo divide la muestra de manera sucesiva por tamaño del hogar, ignorando el resto de las variables. Dado que en la EPH no se cuenta con un conjunto de variables intergeneracionales como en la base utilizada por Brunori et al., prescindir de las variables género y tamaño del hogar dejaría a este estudio prácticamente sin circunstancias relevantes.

En conclusión, dados los problemas mencionados y las variables disponibles en la EPH, se consideró que el ingreso total individual es la variable más apropiada para estimar de la desigualdad de oportunidades en Argentina. En la siguiente sección se explicará la selección de las circunstancias utilizadas para estimar el ingreso total individual.

5.2. La elección de las circunstancias.

Como se explicó en el Capítulo 3, se entiende por circunstancias a aquellas variables que están fuera del control individual. Como ya fue advertido, esta clasificación es un tanto arbitraria, ya que algunas variables pueden estar sólo parcialmente fuera del control individual. De las circunstancias consideradas en este trabajo, hay al menos tres sobre las que considero que no hay discusión respecto a si deben ser incluidas o no:

- **Género**. Aquí no hay mucho que justificar, ya que una persona no elige el sexo con el que nace. También hay bastante consenso respecto a que hay desigualdad salarial en base al género en Argentina (Pérez, 2008; Faur, 2008), siendo las mujeres las perjudicadas.
- **Lugar de nacimiento**. Nuevamente, esta no es una elección que toma el individuo. En particular, al igual que para el caso del género, las diferencias de ingresos explicadas por esta variable pueden ser atribuidas, al menos

parcialmente, a casos de discriminación. Para el caso argentino, existe evidencia respecto a diferencias salariales entre población nativa y migrantes, tal como fue analizado por Maurizio (2008).

- **Nivel educativo de los padres**. Si bien esta variable no se encuentra directamente en la EPH, se puede obtener para aquellas personas que habitan el mismo hogar que al menos uno de sus padres[19]. Es una de las variables más relevantes a la hora de evaluar la desigualdad de oportunidades ya que permite analizar hasta que punto una persona puede revertir una situación inicial desfavorable. Por otro lado, es interesante resaltar que la literatura considera al nivel educativo propio como un esfuerzo del individuo. Esto es discutible, ya que el nivel educativo que alcanza una persona no sólo depende de su esfuerzo individual, sino que está limitado por las circunstancias de su entorno familiar. De todas formas, en este trabajo se decidió no contradecir a la literatura y no incluir el nivel educativo del individuo dentro del conjunto de las circunstancias. Sólo se incluye el de sus padres, en la medida que los datos lo permiten.

Más allá de estas tres circunstancias sobre las que no debería haber demasiada discusión, se incluyen otras variables donde es opinable hasta qué punto están dentro o fuera del control del individuo. Se decidió incluirlas ya sea porque son incluidas en la literatura de referencia o porque no son variables que per sé habiliten una desigualdad de ingresos. Es decir, la parte de la desigualdad de ingresos explicada por estas variables puede considerarse injusta. Las variables en cuestión son las siguientes:

- **Lugar de residencia 5 años atrás.** Puede argumentarse que una persona adulta tiene la libertad de mudarse al lugar que quiere, por lo que el lugar de residencia 5 años atrás pudo haber sido una elección del individuo. Sin embargo, esta supuesta elección puede estar restringida por factores tales como: el lugar de nacimiento, la edad del individuo, oportunidades laborales y/o educativas, situaciones familiares específicas, etc.
- **Cantidad de miembros del hogar.** Hasta cierto punto, una persona puede elegir si forma una pareja o no y cuántos hijos tener. Pero estas decisiones suelen estar limitadas tanto por factores culturales como económicos. Más aún, la presencia

[19] Se filtró a los hogares donde habita al menos uno de los padres del jefe de hogar, y el nivel educativo de esta persona (o el máximo si los dos padres están presentes) se asignó al jefe de hogar y a los hermanos que habitan en el mismo hogar. Del mismo modo con los hogares donde habita al menos uno de los suegros del jefe de hogar. En este caso, el nivel educativo se asignó al cónyuge del jefe de hogar.

en el hogar de padres, suegros, hermanos, cuñados, etc., puede responder más a una adaptación ante un hecho azaroso que a una decisión individual.

- **Cantidad de menores de 10 años en el hogar.** Al igual que para la variable anterior, los individuos tienen un control parcial sobre la cantidad de hijos que tienen, decisión que suele estar influenciada por factores culturales, económicos y fortuitos.
- **Cantidad de trabajadores en el hogar.** Para esta variable se cuenta tanto a los ocupados como a los desocupados, dado que buscan trabajo. En esta variable también hay parte de elección individual, pero depende también de la composición del hogar, de las elecciones de otros miembros del hogar y de factores exógenos que impulsan a determinados miembros del hogar a trabajar o a dejar de hacerlo.
- **Régimen de tenencia de la propiedad.** Para la mayor parte de la población, esto es una circunstancia. La elección entre, por ejemplo, alquilar o comprar, parece estar reservada sólo a sectores privilegiados.
- **Habita o no en una ciudad de más de 500 mil habitantes.** Como ya se mencionó antes, una persona tiene cierta libertad para elegir dónde vivir. Pero muchas veces esta decisión está restringida por motivos económicos, laborales o relacionada con el lugar de nacimiento.

A lo largo de las distintas estimaciones fue necesario realizar algunas decisiones metodológicas respecto a la base datos utilizada. Estas decisiones serán explicadas a medida que se presenten los resultados de la investigación, lo cual se realiza en el siguiente capítulo.

6. Desigualdad de oportunidades en Argentina

En este capítulo se presenta la evidencia empírica destinada a probar las hipótesis formuladas en el Capítulo 2. Los resultados se presentan en tres secciones, cada una de las cuales tiene un objetivo concreto. En la sección primera se muestra que, mediante algoritmos de ML, se pueden obtener estimaciones del nivel de ingresos más precisas que las de metodologías tradicionales, posibilitando una mejor estimación del nivel de desigualdad de oportunidades. Es decir, el objetivo de esta sección es validar el método utilizado en las secciones siguientes para estimar la desigualdad de oportunidades en Argentina. El objetivo

de la segunda sección es probar la relevancia de la única variable intergeneracional que se utiliza: el máximo nivel educativo alcanzado por los padres del individuo. Si bien sería deseable considerar esta variable siempre que se analice la desigualdad de oportunidades, obliga a reducir considerablemente la muestra, afectando la confianza en los resultados. En la tercera sección se abandona la utilización de esta variable, permitiendo trabajar con una base de datos mucho más grande. El objetivo de esta sección será analizar con más detalle el rol y la relevancia de las demás variables incluidas en las estimaciones. Más allá de estos objetivos, tanto en la sección segunda como en la tercera se mostrará evidencia del rechazo de la hipótesis nula de existencia de igualdad de oportunidades. En cuanto a los niveles de desigualdad de oportunidades, a nivel país será estimado en la segunda sección y a nivel regional en la tercera. El capítulo finaliza con una cuarta sección en la que se presentan algunas conclusiones sobre los resultados obtenidos.

6.1. ML vs. metodologías de estimación tradicionales.

Como se explicó en el Capítulo 3, el cálculo del nivel de desigualdad de oportunidades involucra la estimación del nivel de ingresos en base a las circunstancias seleccionadas. A esta estimación se la denomina distribución de ingresos contrafactual y es la parte de la desigualdad que se puede considerar injusta. De este modo, la calidad de la estimación de la distribución de ingresos contrafactual repercute directamente en la calidad de la estimación de la desigualdad de oportunidades. Como también fue explicado, en la literatura se encuentran básicamente dos metodologías para estimar la distribución de ingresos contrafactual, las cuales en este trabajo se han denominado 'paramétrica' y 'no paramétrica'. En esta sección se intenta mostrar que mediante bosques aleatorios condicionales se pueden superar los resultados de estas dos metodologías.

En teoría, los métodos de árboles y bosques son superiores a las metodologías tradicionales cuando hay muchas variables (P) en relación a las observaciones (N), es decir, cuando se tiene un bajo ratio N/P. Es por eso que se experimentará con distintas muestras y cantidad de variables, dando lugar a distintos valores de este ratio. Con cada una de estas muestras se realizarán las siguientes tareas:

1. La muestra se divide en un set de entrenamiento y uno de prueba, en proporciones 2/3 y 1/3 respectivamente.

2. A partir del set de prueba se construyen 200 sets mediante la técnica de *bootstrapping* (extracción con reposición desde el set de prueba, de una cantidad de elementos igual a las del set de prueba). De esta manera, se obtienen 200 sets de prueba que, en principio, pueden ser todos levemente diferentes.
3. Utilizando el set de entrenamiento, se estima un modelo que predice el valor del ingreso total individual en base a las circunstancias consideradas. Esta estimación se realiza mediante cada uno de los tres métodos: paramétrico, no paramétrico y bosque de regresión condicional.
4. Cada uno de los tres modelos estimados en el punto anterior es puesto a prueba realizando predicciones fuera de la muestra para el ingreso total individual en los 200 sets construidos en el paso 2. En cada caso, el valor predicho se compara con el verdadero nivel de ingresos y se calcula la raíz del error cuadrático medio (RMSE, por sus siglas en inglés). De esta manera, para cada una de las metodologías se cuenta con 200 errores de pronóstico.
5. Se comparan los resultados para ver qué metodología se desempeñó mejor.

A continuación, se analizan los resultados de tres casos distintos, los cuales presentan distintos valores del ratio N/P.

6.1.1. Caso 1: Total del país, incluyendo variable intergeneracional, todos los trimestres juntos.

Como ya fue explicado, incluir el nivel de educación de los padres implica reducir la muestra considerablemente: entre los 10 trimestres, se cuentan con tan solo 1.957 observaciones en total. Por lo tanto, realizar una estimación por trimestre y por región hubiera obligado a limitar la cantidad de variables o la cantidad de valores que las mismas pueden obtener. De esta forma, se consideró el total del país, incorporando la variable 'región' como una circunstancia más y se ajustaron los niveles de ingreso de cada trimestre en función a la variación del ingreso promedio en el período. La Tabla 4 resume el factor de ajuste utilizado para llevar el ingreso de cada trimestre a valores del tercer trimestre de 2018.

Para predecir el ingreso se utilizaron 10 variables explicativas[20], tres de las cuales son numéricas y las siete restantes categóricas. Para realizar la estimación mediante el

[20] Región (6), Género (2), Lugar de Nacimiento (5), Nivel educativo del padre/madre (7), Residencia 5 años atrás (5), N° de trabajadores en el hogar, Régimen de tenencia de la propiedad (9), Tamaño del hogar, N° de menores de 10 años en el hogar y variable binaria que indica si se vive en una ciudad de más 500 mil

método paramétrico, se construyeron variables binarias para las variables categóricas[21], siendo requerido estimar 33 parámetros en total. Tomando este valor como P, podemos calcular el ratio $N/P = 59{,}30$.

Tabla 4. Factor de ajuste de los ingresos trimestrales

Trimestre	Ingreso Promedio	Factor de Ajuste
II-2016	$8.961,6	1,922439
III-2016	$10.476,0	1,644523
IV-2016	$10.333,4	1,667216
I-2017	$11.921,5	1,445126
II-2017	$11.776,8	1,462880
III-2017	$13.691,8	1,258276
IV-2017	$13.625,6	1,264389
I-2018	$15.599,5	1,104400
II-2018	$14.845,5	1,160496
III-2018	$17.228,1	1

Para la estimación no paramétrica, es necesario decidir cómo subdividir las variables numéricas. En todos los casos, se decidió dividirlas en dos grupos a partir de la mediana. Dada la cantidad de valores que pueden tomar las variables categóricas, esta metodología implica la construcción de 302.400 tipos de circunstancias[22], lo que implica que la mayoría de estos tipos no contará con ninguna observación. Esto muestra las limitaciones de esta metodología cuando el número de variables es grande respecto a la cantidad de observaciones. En los casos en los que no se cuentan con observaciones de determinado tipo, el nivel de ingresos predicho para ese tipo es igual al ingreso promedio de toda la muestra. Para los tipos que sí cuentan con observaciones, el valor predicho es el ingreso promedio de las observaciones de ese tipo presentes en el set de entrenamiento.

Por su parte, la estimación del bosque aleatorio condicional se realizó mediante el ensamble de 500 árboles de regresión condicionales, mientras que la cantidad de variables consideradas en cada nodo se limitó a 3. Respecto al nivel de significatividad requerido para rechazar o no la hipótesis nula y de esta forma decidir hasta donde se ramifica el árbol, se fijó en un valor exageradamente alto con el fin de producir árboles largos, los cuales suelen arrojar mejores predicciones fuera de la muestra: 90%. En la próxima sección, cuando se

habitantes (2). En el caso de las variables categóricas, se señala entre paréntesis la cantidad de valores que pueden tomar.

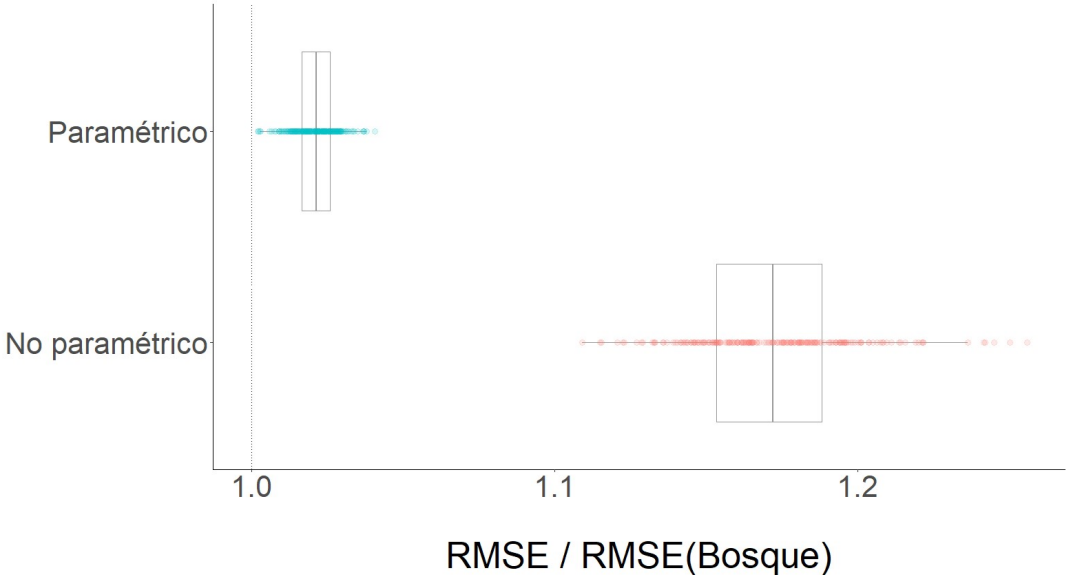
[21] Si la variable categórica tiene n valores posibles, se construyeron $n-1$ variables binarias.

[22] Este resultado surge de multiplicar la cantidad de valores que puede tomar cada una de las 10 variables: 6*2*5*7*5*2*9*2*2*2.

quiera rechazar la hipótesis nula de igualdad de oportunidades, se trabajará con un nivel de la significatividad mucho más exigente: 1% o 10%, dependiendo del tamaño de la muestra.

Como fue explicado, la capacidad predictiva de cada uno de estos modelos fue puesta a prueba en los 200 sets construidos mediante *bootstrapping*. En el siguiente gráfico se muestran los boxplots de los 200 errores de pronósticos cometidos por los métodos parámetrico y no paramétrico, expresados como ratio del error de pronóstico del bosque aleatorio condicional en la misma muestra. Es decir, un valor del ratio mayor a 1 significa que el método en cuestión cometió un error de pronóstico mayor al bosque aleatorio en la misma muestra.

Gráfico 6. Errores de pronóstico relativos al bosque aleatorio. Caso 1.

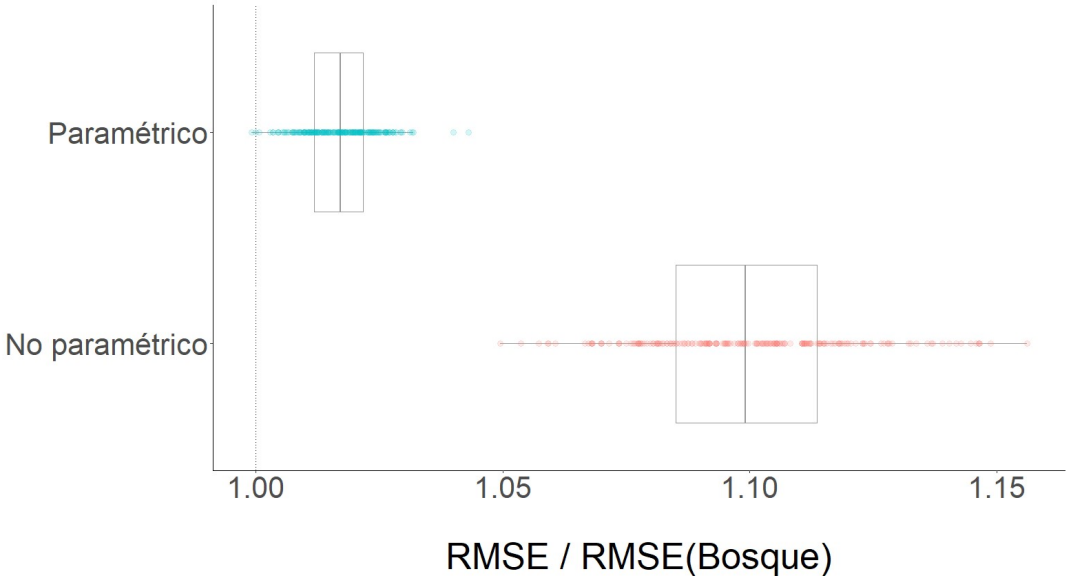


Puede observarse que en el total de las 200 muestras donde se probó la precisión de los modelos, el bosque aleatorio condicional obtuvo un mejor desempeño respecto a las dos metodologías alternativas. Adicionalmente, puede observarse el pobre desempeño del método no paramétrico respecto al método paramétrico. Mientras que el método paramétrico comete un error de pronóstico a lo sumo un 5% superior al del bosque aleatorio, los errores de pronóstico del método no paramétrico pueden ser hasta un 20% superiores. Estos resultados van en línea con lo esperado. A continuación, se probará con ratios N/P mayores, lo que debería reducir las ventajas del bosque aleatorio.

6.1.2. Caso 2: Total del país, incluyendo variable intergeneracional, todos los trimestres juntos, sin considerar todas las circunstancias.

El presente caso es similar al anterior, pero se descartaron algunas circunstancias que, a priori, parecen menos relevantes: lugar de residencia 5 años atrás, régimen de tenencia de la propiedad y la variable binaria que indica si se habita en una región de más de 500 mil habitantes. Como resultado de esto, la cantidad de parámetros a estimar mediante el método paramétrico se redujo a 20, dando lugar a un ratio $N/P = 97{,}85$. Por su parte, al utilizar el método no paramétrico, es necesario construir "apenas" 3360 tipos[23], lo cual sigue siendo superior a la cantidad de observaciones. En todo lo demás, sigue siendo válido lo explicado para el Caso 1. A continuación, se muestran los errores de pronóstico relativos al bosque aleatorio:

Gráfico 7. Errores de pronóstico relativos al bosque aleatorio. Caso 2.



Puede ser difícil de apreciar en el gráfico, pero en 2 de las 200 muestras utilizadas para evaluar la precisión de los modelos, el método paramétrico obtuvo un mejor desempeño que el bosque aleatorio. Aun así, el hecho de que en 198 de las 200 muestras el bosque aleatorio tenga un mejor desempeño muestra claramente la superioridad de este método para esta modelización en particular. Al igual que en el caso anterior, los errores de pronóstico del método paramétrico son, a lo sumo, un 5% superiores a los del bosque aleatorio. Si bien los errores del método no paramétrico siguen siendo mayores, ahora son tan sólo entre un

[23] Este resultado surge de multiplicar la cantidad de valores que puede tomar cada una de las 7 variables: 6*2*5*7*2*2*2. Recordar que para las variables numéricas se consideran dos valores posibles en función de la mediana.

5% y 15% superiores a los del bosque. El último caso considerado en esta sección, se trabaja con ratios N/P que permiten suponer un desempeño más parejo entre las tres metodologías.

6.1.3. Caso 3: Estimaciones por región para cada uno de los trimestres, con sólo algunas de las circunstancias.

En el último caso que se analiza, se descarta la utilización del nivel educativo de los padres, lo que permite trabajar con una muestra significativamente más grande. En particular, se trabajarán con 60 muestras, una por trimestre para cada región, las cuales tienen distinto tamaño, con un rango que va desde 1.488 hasta 4.428 observaciones, tal como se muestra en la siguiente Tabla.

Tabla 5. Estadísticas descriptivas del tamaño de las 60 muestras consideradas

Mínimo	1er cuartil	Mediana	Media	3er cuartil	Máximo
1.488	1.602	2.176	2.532	3.518	4.428

Al trabajar por región, desaparece la variable 'región' como circunstancia. Adicionalmente, tal como se había procedido para el Caso2, se descartaron las variables que, en principio, parecen menos relevantes: lugar de residencia 5 años atrás, régimen de tenencia de la propiedad y la variable binaria que indica si se habita en una región de más de 500 mil habitantes. En concreto, el nivel de ingresos será estimado tan sólo con cinco circunstancias: género, lugar de nacimiento, tamaño del hogar, cantidad de menores de 10 años en el hogar y cantidad de trabajadores en el hogar, donde sólo las dos primeras son categóricas con 2 y 5 valores posibles respectivamente. De esta manera, el método paramétrico sólo requiere estimar 9 parámetros, dando lugar a ratios N/P de entre 165,33 y 492, dependiendo de la muestra en particular. Por su parte, el método no paramétrico requiere la construcción de sólo 80 tipos[24], una cantidad bastante más razonable que en la de los casos previos. La estimación mediante el bosque aleatorio condicional se realizó con los mismos parámetros explicados para el Caso 1.

A continuación, se muestran los gráficos de los errores relativos para cada una de las 60 muestras, agrupados por región.

[24] Este resultado surge de multiplicar los valores posibles de cada variable: 2*5*2*2*2. Recordar que a las variables numéricas se las divide en dos grupos a partir de la mediana.

Gráfico 8. Errores de pronóstico relativos al bosque aleatorio. Caso 3, GBA

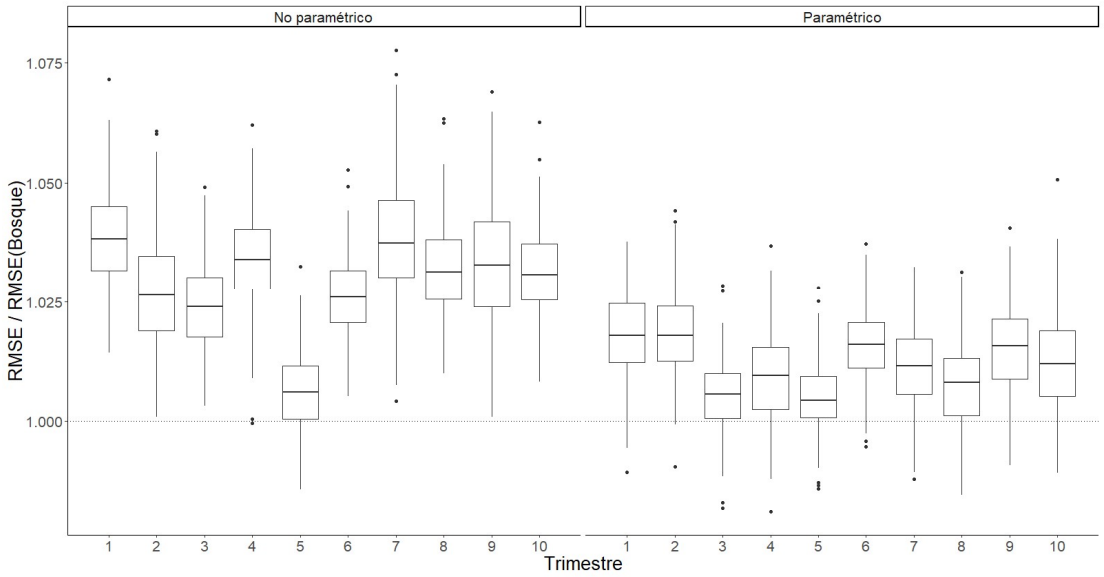


Gráfico 9. Errores de pronóstico relativos al bosque aleatorio. Caso 3, NOA

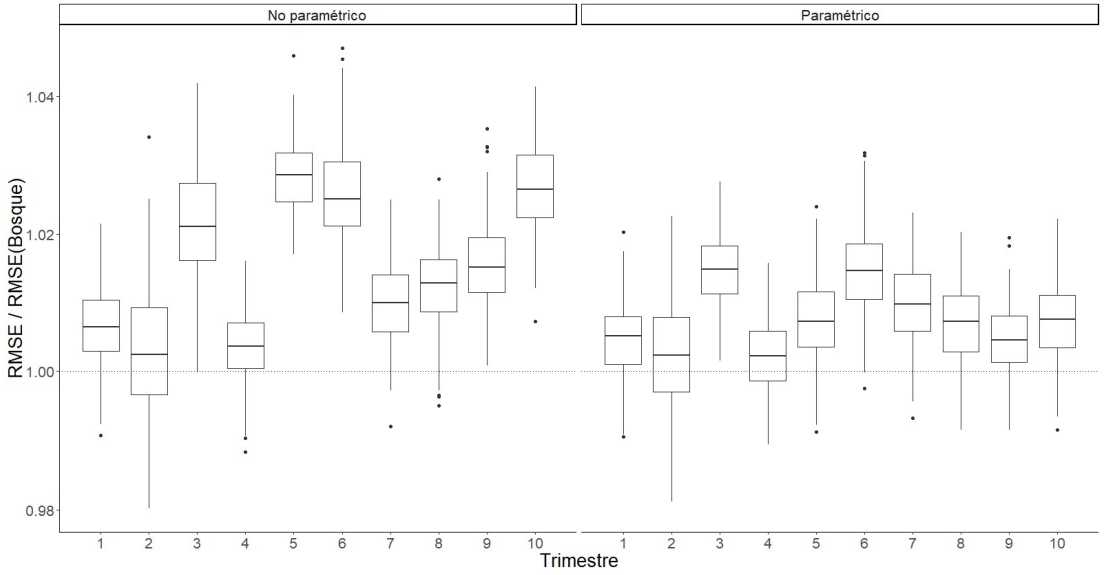


Gráfico 10. Errores de pronóstico relativos al bosque aleatorio. Caso 3, NEA

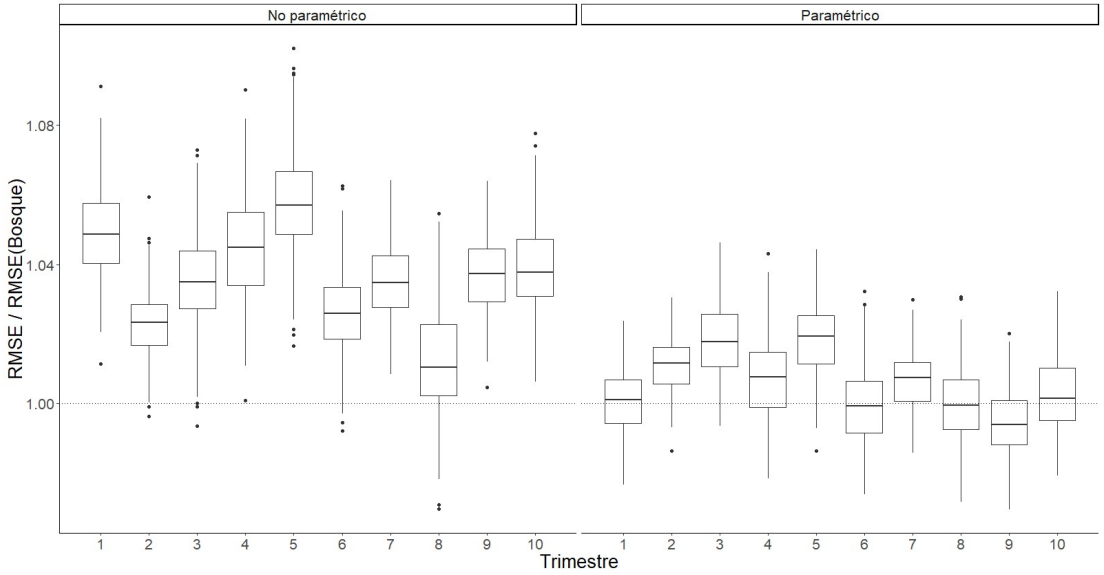


Gráfico 11. Errores de pronóstico relativos al bosque aleatorio. Caso 3, Cuyo

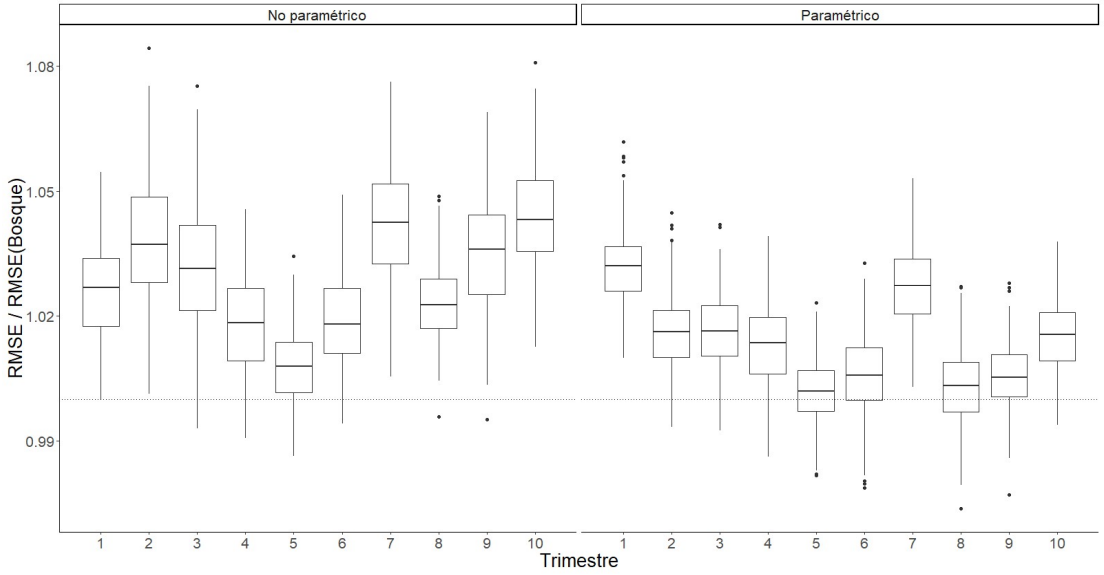


Gráfico 12. Errores de pronóstico relativos al bosque aleatorio. Caso 3, Pampeana

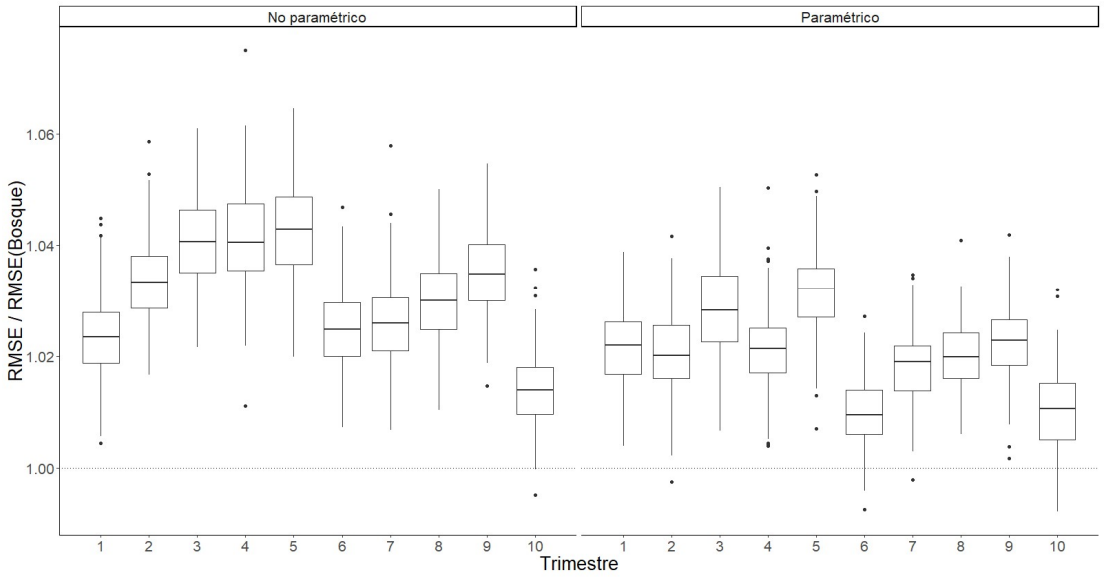
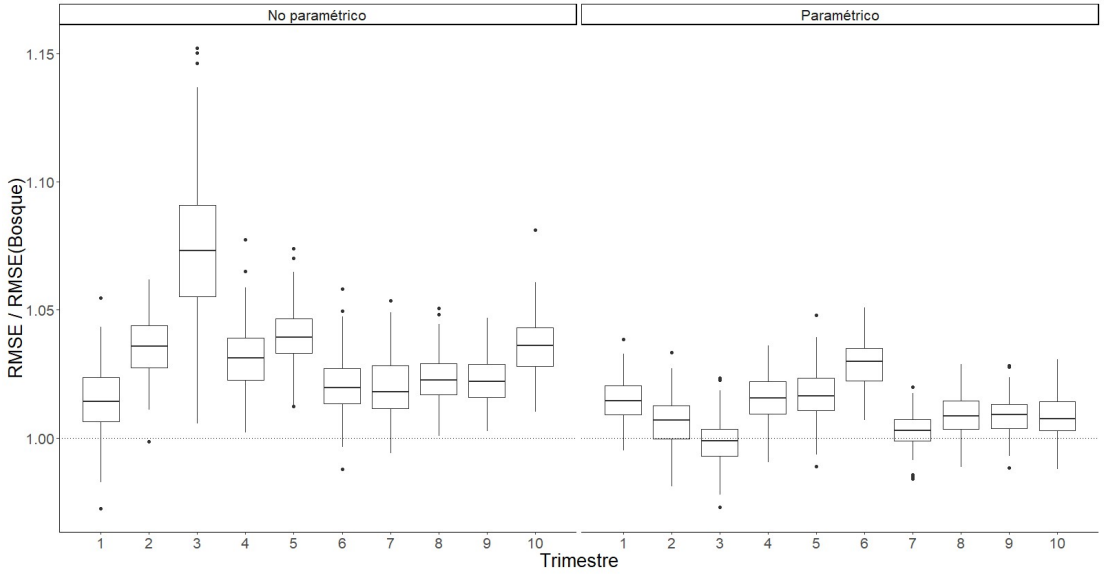


Gráfico 13. Errores de pronóstico relativos al bosque aleatorio. Caso 3, Patagonia



Dos conclusiones saltan a la vista en los gráficos anteriores. En primer lugar, la superioridad de los bosques aleatorios no es tan contundente como en los Casos 1 y 2. En segundo lugar, el método no paramétrico mejora notablemente su performance. Para facilitar la comparación entre las metodologías, aprovechando la información brindada por los boxplots (cuartiles y mediana), se construye la siguiente tabla en la que se resume para cada una de las 60 muestras, en cuántos de los 200 testeos el bosque se desempeño mejor que las otras dos metodologías.

Tabla 6. Análisis del desempeño relativo de las tres metodologías. Porcentaje de casos.

	No Paramétrico	Paramétrico
El bosque es superior en el 100% de los casos	60%	16,67%
El bosque es superior en al menos el 75% de los casos, pero menos que 100%	38,33%	60%
El bosque es superior en al menos el 50% de los casos, pero menos que 75%	1,66%	16,67%
El bosque es superior en al menos el 25% de los casos, pero menos que 50%	0%	6,67%
El bosque es superior en menos del 25% de los casos	0%	0%

De la tabla anterior puede concluirse que, en general, el bosque tiene un desempeño superior a ambas metodologías. Para ninguno de las 60 muestras se encontró que el método no paramétrico comete un error de pronóstico inferior al bosque en la mayoría de los 200 testeos. Por su parte, sólo en 4 de las 60 muestras el método paramétrico comete un error de pronóstico inferior al del bosque en una mayoría de los 200 testeos realizados. Es decir, puede considerarse que los casos en los que el método paramétrico se desempeña mejor que el bosque son excepcionales.

Con el análisis presentado en esta sección, considero validado el algoritmo de los bosques aleatorios condicionales como método para estimar la desigualdad de oportunidades. Permite obtener estimaciones más confiables de la distribución del ingreso contrafactual que los métodos paramétricos y no paramétricos que se encuentran en la literatura. Este desempeño superior es muy claro cuando el tamaño de la muestra es chico comparado con la cantidad de circunstancias a considerar, como es el caso cuando se desea incluir el nivel educativo de los padres en el análisis. Cuando el ratio entre observaciones y parámetros no es tan alto, el desempeño de las metodologías es más parejo, pero aún así el bosque aleatorio se desempeña mejor en la mayoría de los casos.

En la siguiente sección, comienza el análisis de la desigualdad en Argentina mediante árboles y bosques aleatorios condicionales.

6.2. Desigualdad de oportunidades considerando una variable intergeneracional

Como ya fue señalado, el objetivo central de esta sección es probar la relevancia de la única variable intergeneracional considerada para explicar la desigualdad de oportunidades. Adicionalmente, se realizarán distintos testeos de la hipótesis nula de existencia de igualdad de oportunidades y se presentará una estimación numérica a nivel país del grado de desigualdad de oportunidades.

El análisis se hará en distintos niveles. En primer lugar, se trabajará a nivel país con una base similar a la del Caso 1 de la sección anterior. En segundo lugar, se presentarán resultados a nivel regional. La primera aproximación en cada caso será a través de los árboles de regresión condicionales, dada la sencilla interpretación de sus resultados. Este análisis será seguido de los resultados estimados mediante bosques aleatorios condicionales, que si bien son menos transparentes, permiten una estimación más precisa del nivel de desigualdad y de la importancia de cada variable. Esto se debe a que, como fue explicado en el Capítulo 4, las estimaciones mediante árboles tienen gran varianza, la cual se reduce con el ensamble de árboles que da lugar al bosque.

Luego de filtrar la base para aquellos individuos de los que se puede saber el nivel educativo de sus padres y de juntar todos los trimestres ajustando el nivel de ingresos tal como se explicó en la sección anterior, quedan 1.957 observaciones distribuidas de la siguiente manera:

Tabla 7. Cantidad de observaciones considerando la variable intergeneracional

REGIÓN	OBSERVACIONES
Gran Buenos Aires	447
NOA	459
NEA	183
Cuyo	203
Pampeana	472
Patagonia	193
Total	**1.957**

El análisis a nivel país se hará considerando las siguientes 10 circunstancias: Región, Género, Lugar de Nacimiento, Máximo nivel educativo alcanzado por alguno de sus padres, Lugar de residencia hace 5 años, Régimen de tenencia de la propiedad, Cantidad de miembros del hogar, Cantidad de menores de 10 años en el hogar, Cantidad de trabajadores en el hogar y una variable binaria que indica si se habita en una ciudad de más de 500 mil

habitantes. El análisis a nivel regional se hará con las mismas circunstancias, con la obvia excepción de la variable 'región'. Para interpretar los gráficos, tener en cuenta la codificación de los valores que puede tomar cada variable, disponible en la Tabla 3.

6.2.1. Total del país

Tanto para los árboles de regresión como para los bosques aleatorios se trabajó con un nivel de significatividad del 1%. Es decir, la división de cada árbol se produce sólo cuando se puede rechazar la hipótesis nula de existencia de igualdad de oportunidades con una confianza del 99%. En el Gráfico 14 se presentan los resultados del árbol de regresión.

La forma de interpretar el árbol es la siguiente. En primer lugar, el hecho de que el árbol no sea vacío significa que se rechazó con un 1% de significatividad la hipótesis nula de que la distribución de ingresos es independiente de las circunstancias. Es decir, se rechazó la hipótesis nula de la existencia de igualdad de oportunidades. Para esto, en primer lugar el algoritmo rechazó la hipótesis global, procediendo luego a realizar los tests de hipótesis individuales. De esta forma, el algoritmo selecciona la variable y el valor de corte que produce un menor valor-p en el test de hipótesis, utilizándolo para dividir la muestra. En este caso, el algoritmo encontró que separando la muestra entre los individuos que viven en la Patagonia y los que viven en el resto del país se obtiene el menor valor-p al realizar el test de hipótesis[25]. Como se muestra en el gráfico, en este caso el valor-p fue inferior al 0,001. Luego, el algoritmo procede de manera similar para cada rama del árbol. Para los 193 casos que viven en la Patagonia, el algoritmo no consigue volver a rechazar la hipótesis nula, por lo que estaría indicando que, al interior de la Patagonia existe igualdad de oportunidades. Para el resto del país sí se rechaza la hipótesis nula, en primer lugar en base al tamaño del hogar y luego en base al género, nivel educativo de los padres, lugar de nacimiento y, nuevamente, región del país. En total, pueden identificarse 7 grupos de individuos, con diferentes oportunidades entre sí, pero con igualdad de oportunidades en su interior. En la Tabla 8 se resumen los 7 grupos que identifica el algoritmo.

Estos resultados deben tomarse con precaución debido a la ya mencionada volatilidad de las estimaciones mediante árboles de regresión. Un ligero cambio en la muestra, puede dar lugar a resultados muy distintos. De todas formas, sirven para dar una idea de cuáles son

[25] Este resultado puede sonar obvio, surgiendo la crítica de por qué no se controló por región. Al final de la sección se brinda una respuesta a esta potencial crítica.

las variables relevantes para explicar la desigualdad de oportunidades y en qué sentido afectan a la misma.

Gráfico 14. Árbol de Oportunidades de Argentina, con variable intergeneracional

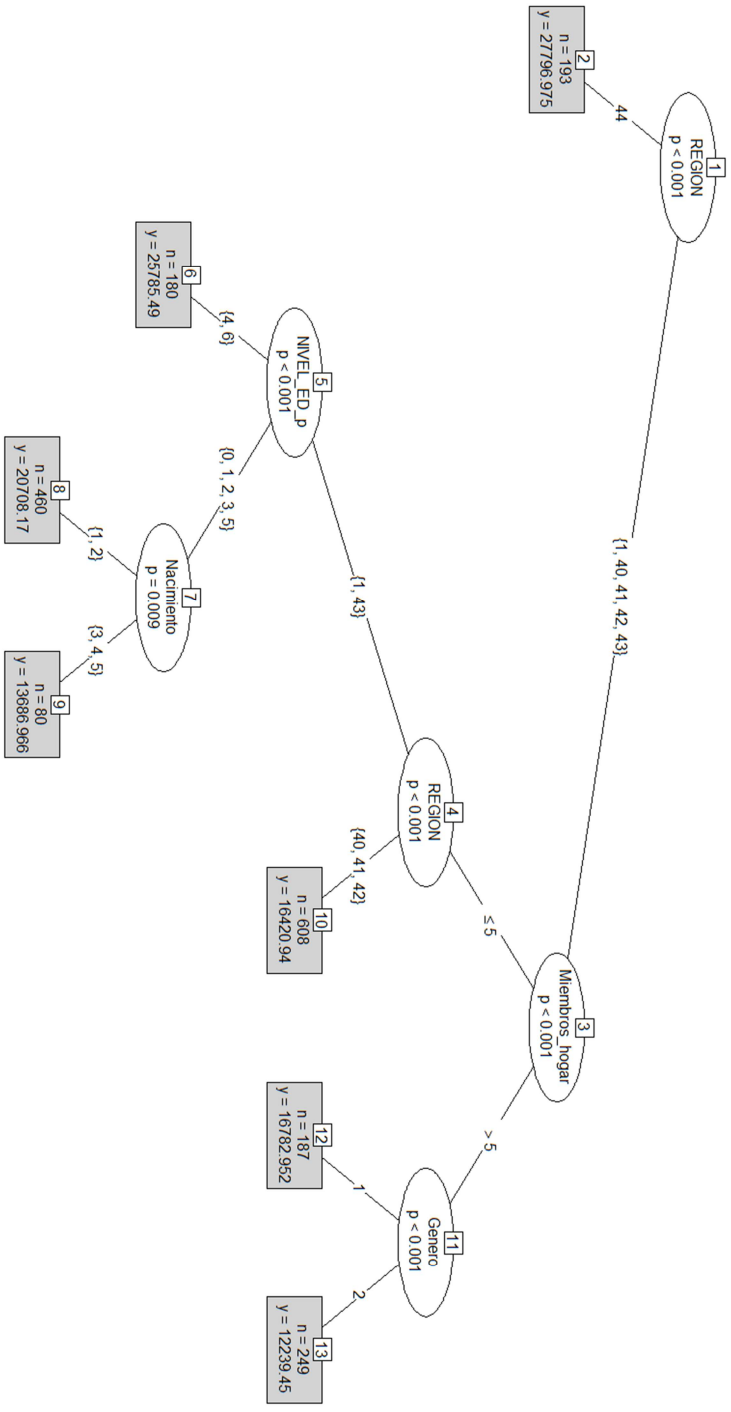


Tabla 8. Grupos de oportunidades de Argentina, con variable intergeneracional

Grupos de Oportunidades	Ingreso Medio
Habitantes de la Patagonia	$27.797
Personas que viven en hogares de hasta 5 miembros, ubicados en el GBA o región Pampeana, con padres que tienen educación secundaria o universitaria completa	$25.785
Personas que viven en hogares de hasta 5 miembros, ubicados en el GBA o región Pampeana, con padres que tienen otro nivel educativo, y que han nacido en la localidad que habitan o en la misma provincia.	$20.708
Personas que viven en hogares de hasta 5 miembros, ubicados en el GBA o región Pampeana, con padres que tienen otro nivel educativo, y que han nacido fuera de la provincia que habitan.	$13.687
Personas que viven en hogares de hasta 5 miembros, ubicados en el NOA, NEA o Cuyo.	$16.421
Hombres en hogares de más de 5 miembros que no habitan en la Patagonia	$16.783
Mujeres en hogares de más de 5 miembros que no habitan en la Patagonia	$12.239

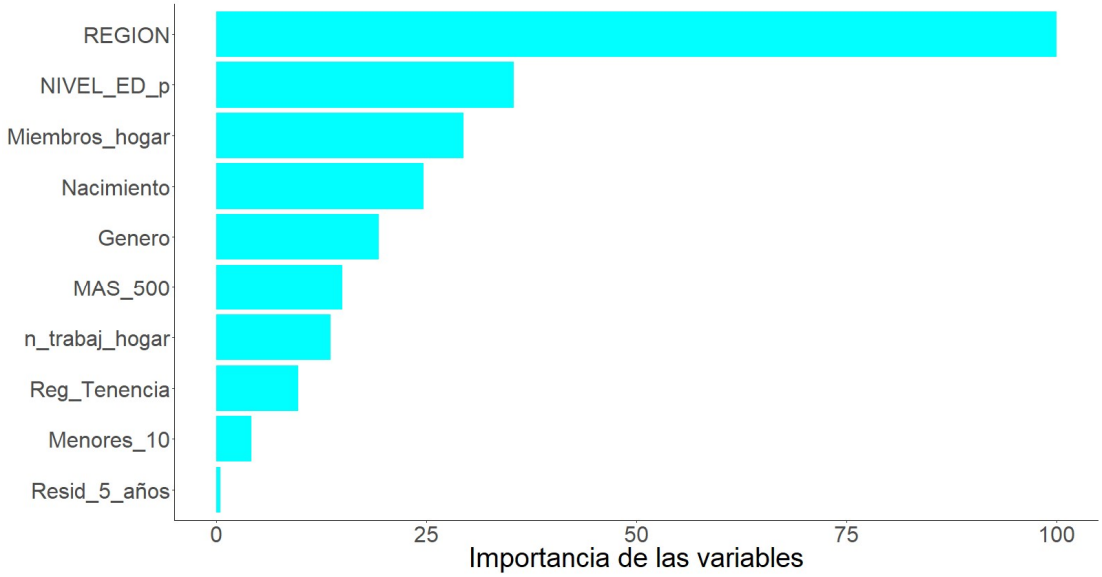
Es por eso que, si bien los bosques aleatorios no permiten una ilustración tan atractiva como la anterior, son muy útiles para realizar una estimación más precisa sobre la importancia de las variables. Tal como fue explicado en el Capítulo 4, el cálculo de la importancia de las variables se realiza mediante predicciones fuera de la muestra[26] y utilizando permutaciones en los valores de las variables para distinguir entre asociaciones reales entre las variables de las que surgen del mero azar. El Gráfico 15 muestra los resultados de la importancia de las variables estimadas para esta muestra, donde los valores se han normalizado para que la variable más importante tenga un valor de 100.

Puede observarse que la región en la que se habita es claramente la variable más importante, resultado esperable y que coincide con lo que había mostrado el árbol. Pero más interesante es observar que la variable que le sigue en importancia es el nivel educativo de los padres, lo cual justifica el intento de incluirla pese a que eso implique reducir considerablemente la muestra. Siguen en importancia la cantidad de miembros del hogar, el lugar de nacimiento y el género. Cabe destacar que las tres variables que de manera indiscutida están fuera del control individual (nivel educativo de los padres, lugar de nacimiento y género) están dentro de las cinco variables más importantes a la hora de

[26] En particular, mediante predicciones *out-of-bag* (OOB), tal como fue explicado en el Capítulo 4.

explicar la distribución del ingreso, lo que refuerza la idea de existencia de igualdad de oportunidades.

Gráfico 15. Importancia de las variables para Argentina, con variable intergeneracional



Por último, mediante los bosques aleatorios se estimó una distribución de ingresos contrafactual, sobre la que se puede aplicar un indicador de desigualdad dando lugar a una medida del nivel de desigualdad de oportunidades. Aplicando el coeficiente Gini a los resultados en esta muestra, se obtiene un valor de 0,1556. Si bien los resultados no son estrictamente comparables, el valor de este coeficiente Gini es aproximadamente el doble que los estimados por Brunori et al. (2018) para Alemania.

Antes de finalizar esta sección, es necesario dar una explicación adicional sobre las diferencias entre los métodos de árboles y las estimaciones paramétricas, de manera de interpretar adecuadamente el análisis que se está llevando a cabo. Una primera lectura de los resultados anteriores puede ser que los mismos están distorsionados por las desigualdades regionales de nuestro país. Los altos ingresos de la región patagónica pueden estar asociados a un mayor costo de vida, por lo que no necesariamente puedan relacionarse con un mejor nivel de vida. Es decir, contar con mayores ingresos en la Patagonia que en el GBA no necesariamente implica tener mejores oportunidades. Esto es cierto, por lo que podría surgir la pregunta de por qué no se aplicó algún tipo de corrección para homogeneizar los ingresos por región, tal como se hizo para combinar los distintos trimestres. Básicamente, la respuesta es que, al igual que una regresión lineal (método paramétrico), el árbol es capaz de lidiar con estos distintos 'ingresos medios por región'.

Si la estimación se hubiera realizado mediante una regresión lineal, incluyendo una variable binaria por región, el modelo hubiera considerado los diferentes ingresos medios regionales a la hora de estimar el rol de las restantes variables. El árbol de regresión también es capaz de lidiar con la variable 'región', aunque de una manera diferente. En particular, sólo se considera esta variable y a determinada realización de la misma cuando es la seleccionada para dividir el árbol. Esto tiene varias consecuencias. En primer lugar, las regiones que se terminan conformando están determinadas por el modelo y los datos, no necesariamente respetando la regionalización dada por la variable utilizada. Por ejemplo, en el árbol del Gráfico 14, el GBA y la región Pampeana se consideran como una sola región, al igual que el NOA, NEA y Cuyo. Es decir, el algoritmo decidió dividir la población en 3 regiones, en lugar de las seis predeterminadas. En segundo lugar, el rol de las restantes variables no se analiza considerando el total de las observaciones, como sí ocurre en la regresión lineal. Esto puede explicar por qué no se rechazó la hipótesis nula de igualdad de oportunidades al interior de la Patagonia: el testeo sólo se realizó considerando las 193 observaciones de esa región, no las 1.957 que componen la muestra. En tercer lugar, aunque muy relacionado con lo anterior, una importante observación resaltada por Strobl et al. (2009): los árboles y bosques son muy buenos para analizar interacciones entre variables, pero no tanto para analizar efectos directos. En efecto, si a partir del Gráfico 14 se quisiera saber cuál es el efecto en los ingresos de vivir en la región pampeana, sería imposible dar una respuesta única, como la que podríamos obtener mediante una estimación paramétrica. Sólo se pueden mencionar el efecto que tiene en hogares de hasta cinco integrantes e interactuando con el nivel educativo de los padres y el lugar de nacimiento. Es por eso que, tal como fue explicado en el Capítulo 4, que este tipo de algoritmo sea o no la elección correcta depende de la pregunta que se quiere responder. En este caso, dado que lo que se prioriza es una buena estimación de la distribución de ingresos y una evaluación de la relevancia de cada variable, son una elección apropiada.

De todas formas, luego de considerar todas las observaciones anteriores, vale la pena realizar estimaciones al interior de cada región del país, aunque esto requiera trabajar con muestras aun más reducidas. Esto se realiza en la siguiente sección.

6.2.2. Nivel regional.

Construir árboles con muestras muy reducidas, tal como se intenta en esta sección, tiene sus problemas. Recordemos que el método consiste en ir particionando recursivamente

la muestra, por lo que, si se parte de un tamaño muestral reducido, en seguida se llega a submuestras con muy pocas observaciones. Es decir, los árboles que se obtendrían tendrían muy poca complejidad o incluso serían vacíos. Para poder realizar un análisis un tanto más rico sobre las distintas regiones, se decidió flexibilizar el nivel de significatividad con el que se testea la hipótesis nula, aunque sólo en el caso de los árboles. En concreto, en los árboles que se muestran a continuación se trabajó con un nivel de significatividad del 10%. Pero para la estimación de la importancia de cada variable mediante bosques aleatorios, se siguió trabajando con un 1% de significatividad.

Gráfico 16. Árbol de Oportunidades del GBA, con variable intergeneracional

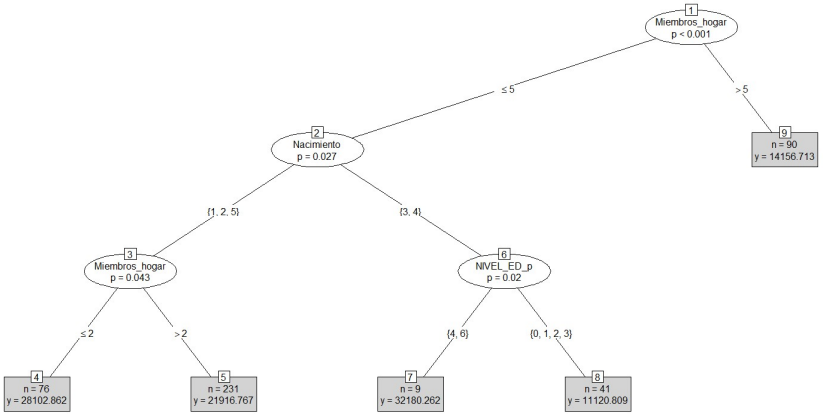


Gráfico 17. Árbol de Oportunidades del NOA, con variable intergeneracional

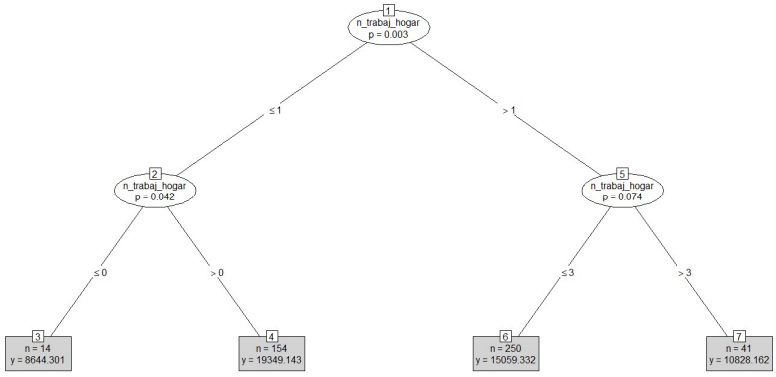


Gráfico 18. Árbol de Oportunidades del NEA, con variable intergeneracional

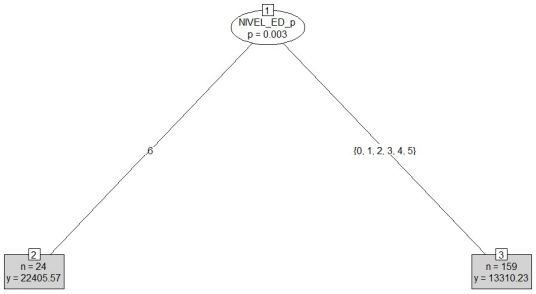


Gráfico 19. Árbol de Oportunidades de Cuyo, con variable intergeneracional

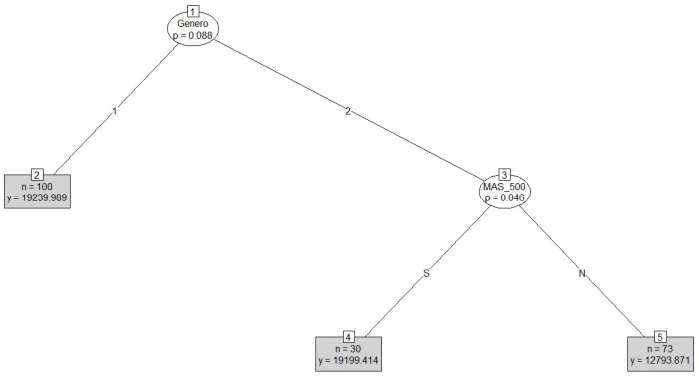


Gráfico 20. Árbol de Oportunidades de Región Pampeana, con variable intergeneracional

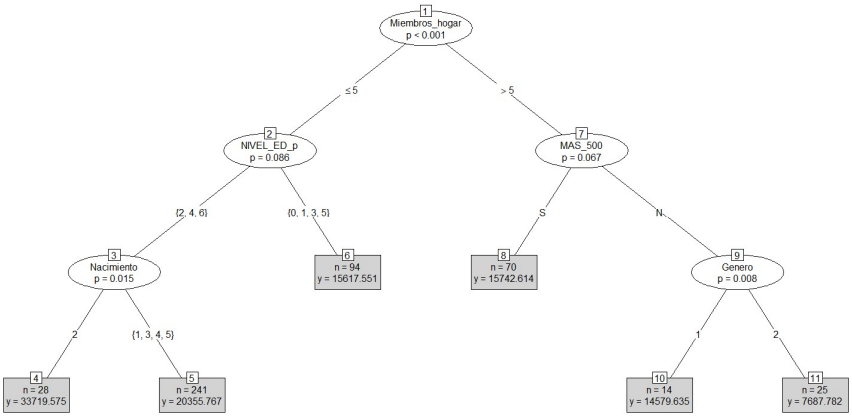
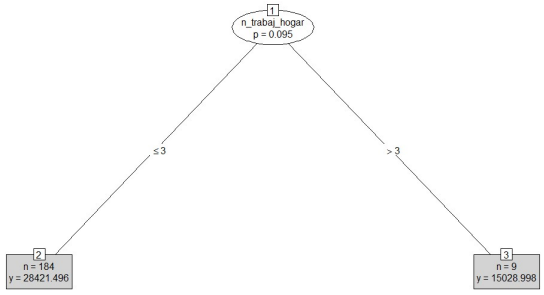


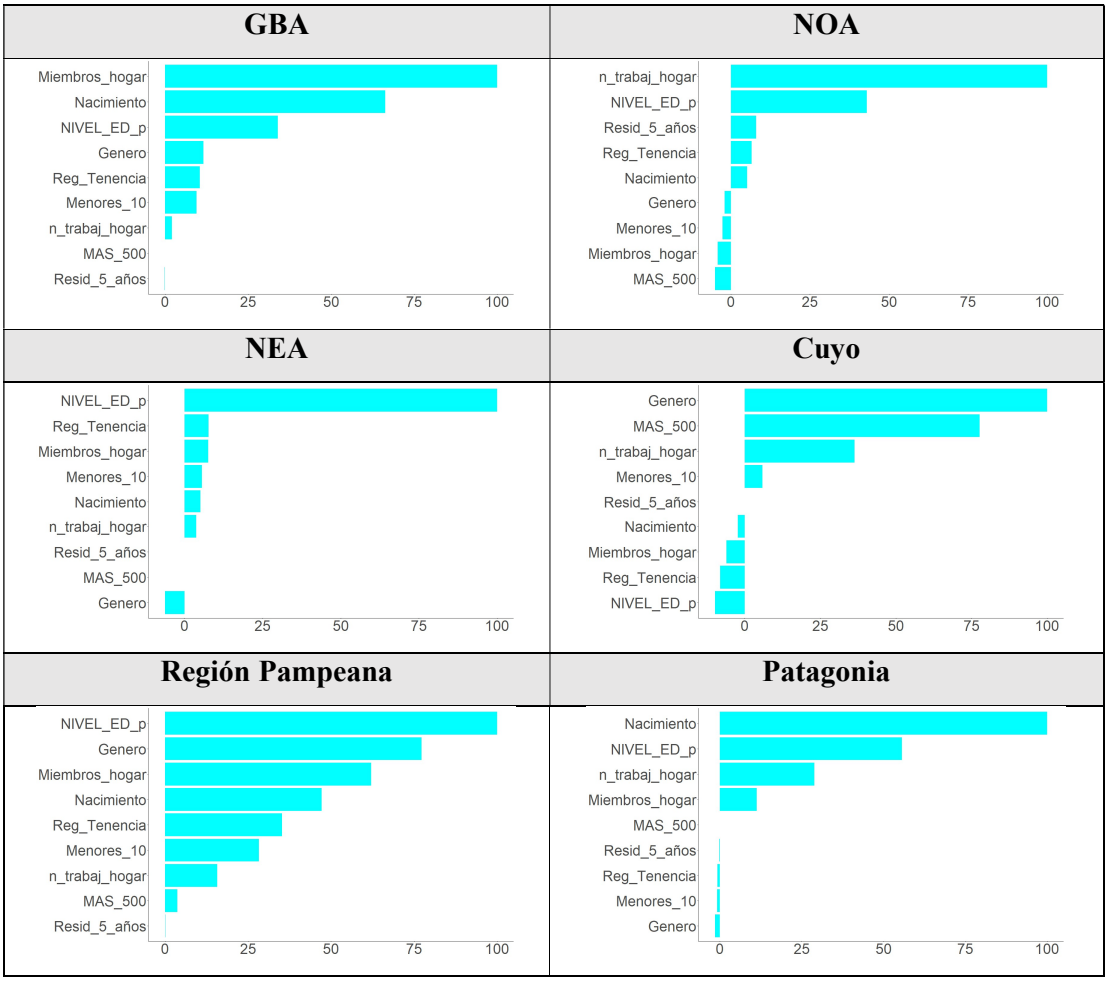
Gráfico 21. Árbol de Oportunidades de la Patagonia, con variable intergeneracional



Lo primero que hay que resaltar de los árboles anteriores es que, en todos los casos, se rechazó la hipótesis nula de existencia de igualdad de oportunidades. Es decir, por más volátiles que puedan resultar los árboles, en todos los casos se encontró evidencia al 10% de significatividad de que la distribución del ingreso no es independiente de las circunstancias. En segundo lugar, en el Gráfico 22 puede apreciarse que el nivel educativo de los padres es la variable más relevante en 2 de las seis regiones (NEA y Pampeana), la segunda en importancia en el NOA y Patagonia y la tercera en importancia en el GBA. Sólo en la región de Cuyo no se encontró evidencia de la relevancia de esta variable para explicar los

ingresos[27]. Es decir, en cinco de las seis regiones del país, el nivel educativo de los padres es relevante para explicar el nivel de ingresos. Por su parte, la variable género es muy relevante tanto en Cuyo como en la Región Pampeana, mientras que el lugar de nacimiento lo es en el GBA y en la Patagonia. El tamaño del hogar y la cantidad de trabajadores también son relevantes en algunas regiones.

Gráfico 22. Importancia de las variables, con variable intergeneracional



Como fue señalado, el objetivo central de esta sección era testear la relevancia de la variable intergeneracional, es decir, del nivel educativo de los padres, para explicar la desigualdad de oportunidades. Considero que este objetivo está cumplido. Más aún, lo que se observa en los árboles en donde aparece esta variable es que, tal como se podría suponer,

[27] Recordar que la medida de importancia de las variables no es un porcentaje, sino que surge de comparar las predicciones que se producen con la variable original respecto a la variable con sus valores permutados. Es por eso que, para variables que no son relevantes, la diferencia puede ser negativa, como en este caso.

un mayor nivel educativo en los padres está asociado con un mayor nivel de ingreso de los hijos. Es decir, existen límites a la movilidad social. Una persona que nace en un hogar con bajo nivel educativo parece tener menos oportunidades para obtener ingresos elevados en su vida adulta.

Para las demás variables, se realiza un análisis más detallado en la sección siguiente, donde se trabaja con una muestra más grande.

6.3. Desigualdad de oportunidades por región y trimestre

En esta sección del trabajo se abandona la utilización de la variable intergeneracional, lo que permite contar con muestras sensiblemente más grandes para estimar la desigualdad de oportunidades a nivel regional y analizar su evolución en el tiempo. El análisis se dividirá en etapas. En primer lugar, se presentarán los resultados de los árboles de regresión condicionales, dada la utilidad que tienen para interpretar los resultados. En particular, sólo se mostrarán los resultados para el último trimestre disponible, el tercero de 2018, presentándose luego un resumen de los resultados en el total de diez trimestres. Luego, se presentan los resultados de la estimación mediante bosques aleatorios condicionales. En particular, se presenta la importancia promedio de cada variable a lo largo de los diez trimestres y la evolución en el tiempo del indicador de desigualdad oportunidades calculado mediante el coeficiente de Gini. Los resultados completos de cada trimestre y región, tanto para el caso de los árboles como de los bosques, están disponibles en el anexo online[28].

6.3.1. Árboles de oportunidad

Utilizando árboles de regresión condicionales, se realizó una estimación por región para cada uno de los diez trimestres disponibles. En todos los casos, el ingreso total individual fue estimado mediante ocho circunstancias: Género, Lugar de Nacimiento, Lugar de residencia hace 5 años, Régimen de tenencia de la propiedad, Cantidad de miembros del hogar, Cantidad de menores de 10 años en el hogar, Cantidad de trabajadores en el hogar y una variable binaria que indica si se habita en una ciudad de más de 500 mil habitantes. Nuevamente, la codificación de los valores que puede tomar cada variable es la misma que la de la Tabla 3. En todos los casos se trabajó con un nivel de significatividad del 1%.

[28] Ver Anexo I para mayor información.

A continuación, se presentan los árboles estimados para el tercer trimestre de 2018, para el cual se dispone de la siguiente cantidad de observaciones:

Tabla 9. Cantidad de observaciones para el 3er. Trimestre de 2018

REGIÓN	OBSERVACIONES
Gran Buenos Aires	2.392
NOA	3.600
NEA	1.496
Cuyo	1.614
Pampeana	4.221
Patagonia	2.030
Total	**15.353**

Nuevamente, la hipótesis nula de existencia de igualdad de oportunidades se rechazó en todos los casos con un 1% de significatividad. Y, en todos los casos, la primera ramificación del árbol se produce en base a la variable género: las mujeres obtienen menos ingresos que los varones. Más aún, en todos los árboles aparece la variable cantidad de miembros del hogar sólo en la rama femenina. Es decir, pareciera ser que las mujeres no sólo pierden por su condición de mujer sino que son las que sufren las consecuencias de habitar en un hogar numeroso. A los hombres, por el contrario, el tamaño del hogar parece no afectarles. Seguramente esto tiene que ver con cómo se ha dado tradicionalmente la división del trabajo entre los géneros, siendo las mujeres quienes cargan en mayor medida con la responsabilidad del cuidado de las personas que habitan el hogar.

Gráfico 23. Árbol de oportunidades GBA, 3er Trimestre de 2018

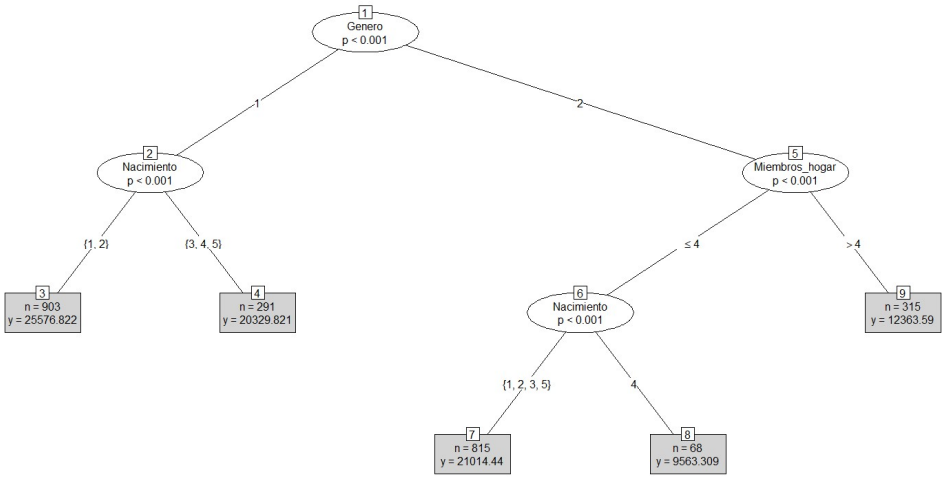


Gráfico 24. Árbol de oportunidades NOA, 3er Trimestre de 2018

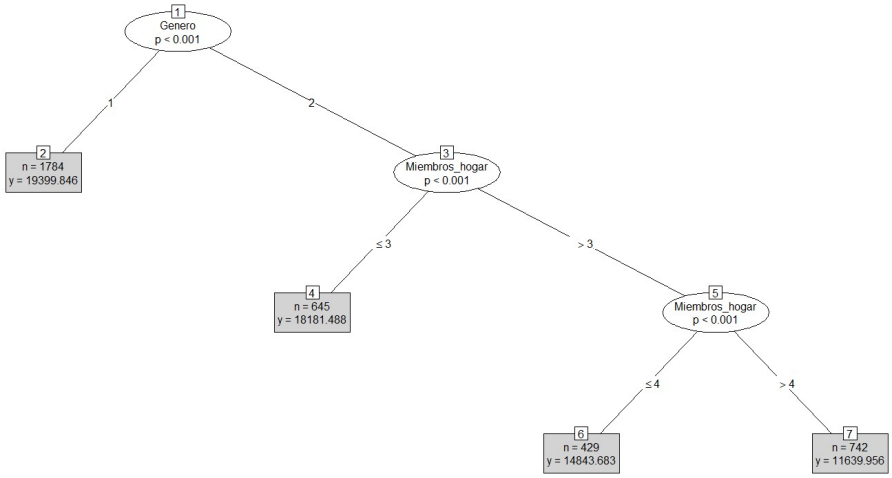


Gráfico 25. Árbol de oportunidades NEA, 3er Trimestre de 2018

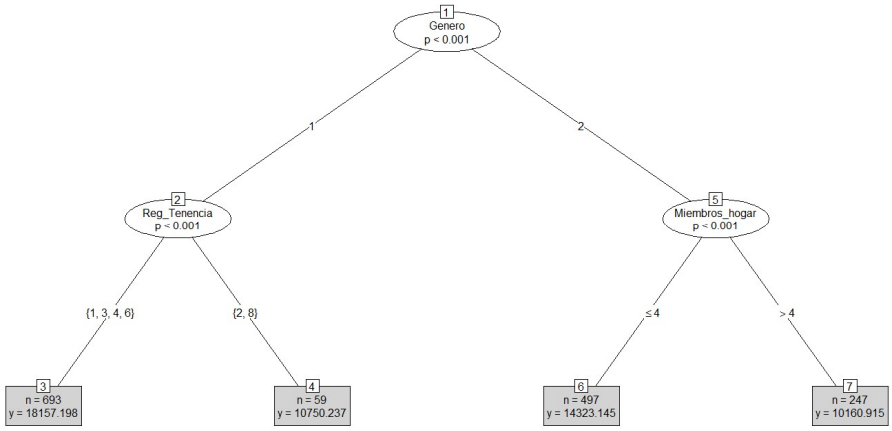


Gráfico 26. Árbol de oportunidades Cuyo, 3er Trimestre de 2018

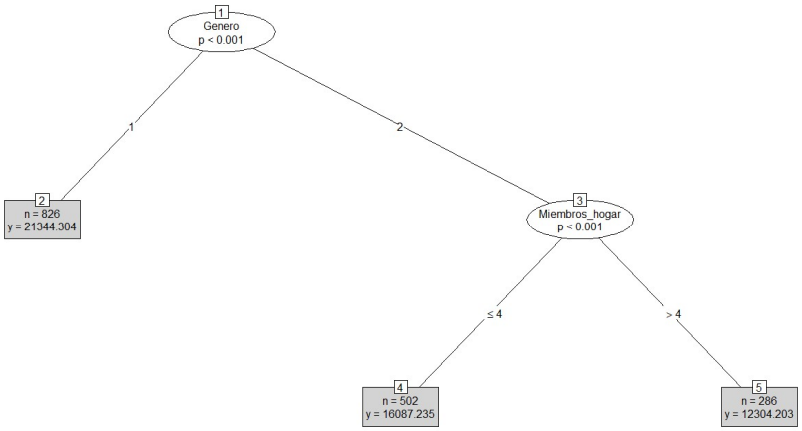


Gráfico 27. Árbol de oportunidades Región Pampeana, 3er Trimestre de 2018

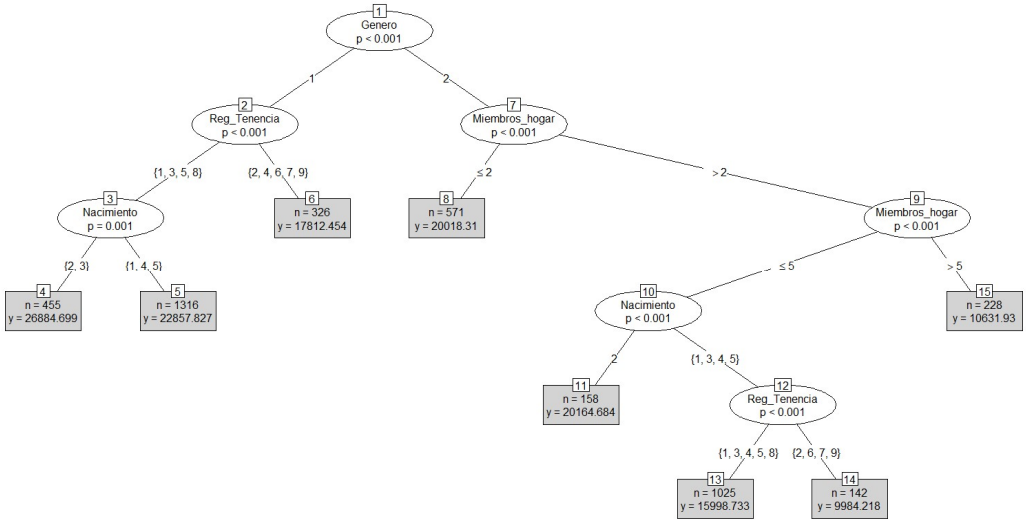
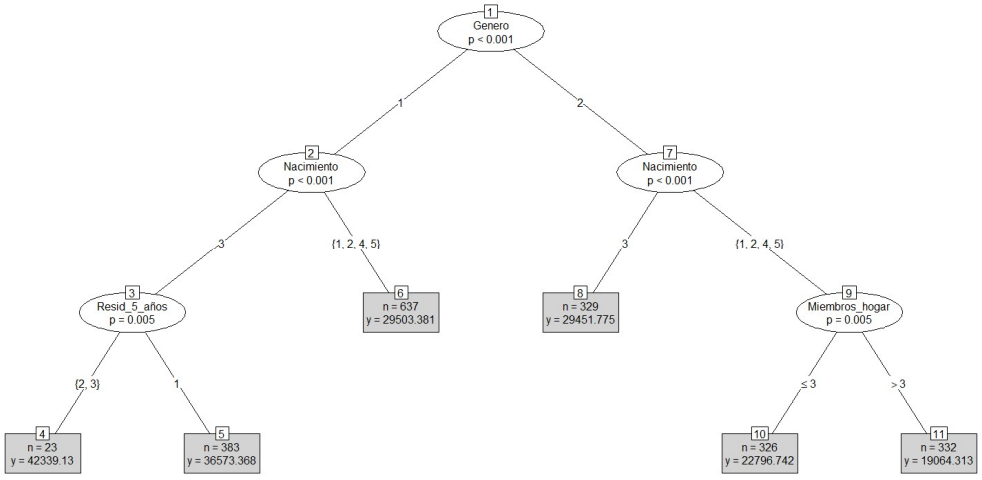


Gráfico 28. Árbol de oportunidades Patagonia, 3er Trimestre de 2018



El lugar de nacimiento parece jugar distinto rol en las distintas regiones. En GBA los hombres nativos tienen mayores ingresos que aquellos que vienen de otras provincias u otros países, mientras que, en el caso de las mujeres, quienes nacieron en países limítrofes tienen menores ingresos que el resto. En la Patagonia, por su parte, los mayores ingresos los obtienen quienes vienen de otras provincias. Más aún, al menos en el caso de los hombres, quienes nacieron en otra provincia pero ya estaban en la localidad hace cinco años, obtienen menos ingresos que los que llegaron más recientemente. Un caso intermedio parece ser la región Pampeana, donde los hombres con mayor ingreso son los que nacieron en otra provincia o en otra localidad de la misma provincia. En el caso de las mujeres de esta región, el mayor ingreso corresponde solamente a quienes nacieron en otra localidad de la misma provincia. Es probable que lo observado para el GBA esté vinculado con las migraciones desde países limítrofes de personas de bajos ingresos, mientras que los patrones observados en las regiones Pampeana y Patagónica estén asociadas a la expansión de ciertas actividades productivas que estén acompañadas de procesos migratorios internos. Sin lugar a duda, este es un resultado que amerita un mayor análisis que excede los objetivos de este trabajo.

La última variable que aparece en dos de los árboles es el régimen de tenencia de la propiedad. En el NEA, los hombres propietarios sólo de la vivienda pero no del terreno y aquellos que se encuentran en sucesión son los que obtienen menores ingresos. En la región Pampeana, por su parte, los mayores ingresos corresponden a los propietarios de vivienda y terreno, los inquilinos, los ocupantes en relación de dependencia y quienes se encuentran en sucesión.

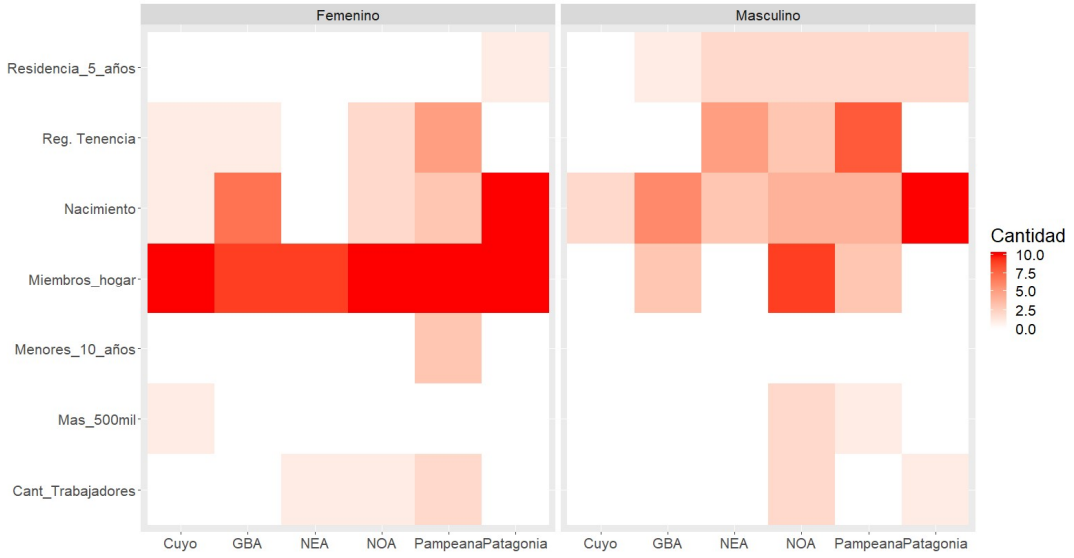
Ya se mencionó que las estimaciones mediante árboles presentan alta varianza, por lo que es importante verificar cuáles de estos resultados se mantienen en los nueve trimestres restantes, los cuales se encuentran disponibles en el anexo online. Analizando los mismos, pueden sacarse las siguientes conclusiones:

- En el 100% de los casos se rechazó la hipótesis nula de existencia de igualdad de oportunidades al 1% de significatividad. Esto se refleja en que ninguno de los 60 árboles resultó vacío.
- La variable género aparece en el 100% de los árboles, sugiriendo siempre un menor ingreso para las mujeres. En 48 de los 60 árboles, es la primer variable utilizada para dividir la muestra. Los 12 casos donde esto no ocurre corresponden al NOA y la Patagonia. En 8 de los 10 árboles del NOA, la primer variable es la cantidad de miembros del hogar, mientras que en 4 de los 10 árboles de la Patagonia, la primer variable es el lugar de Nacimiento. En todos los casos, la variable género se encuentra uno o dos niveles más abajo.
- La mencionada interacción entre la variable género y el tamaño del hogar parece sostenerse. En 58 de los 60 árboles la variable tamaño del hogar aparece en la rama femenina del árbol, mientras que sólo aparece 15 veces en la rama masculina. En todos los casos, un hogar más grande está asociado a menores ingresos.
- El lugar de nacimiento aparece en los 10 árboles del GBA. Parece bastante robusto el resultado de que los nativos obtienen mayor ingreso que los nacidos en otro lugar, siendo particularmente perjudicados los nacidos en países limítrofes. Por el contrario, en los 10 árboles de la Patagonia se sugiere que quienes nacieron en otra provincia obtienen mayores ingresos. Algo similar ocurre en el NOA, NEA y Cuyo, aunque la variable no aparece en todos los árboles. En la región pampeana, por su parte, los nacidos en otra localidad de la misma provincia parecen ser los más favorecidos.

Dado que la variable género aparece en la totalidad de los árboles y, mayormente, en las primeras ramificaciones del árbol, puede ser interesante ver cómo interactúa esta variable con las demás, para cada región. Esto puede observarse en el siguiente mapa de color, donde

un color blanco indica que dicha variable no interactúa con la variable género en ninguno de los 10 árboles, mientras que un color rojo oscuro indica que interactúa en todos los árboles.

Gráfico 29. Mapa de calor con interacción de género y región con las demás variables



Puede observarse claramente cómo la cantidad de miembros del hogar es muy relevante para el caso de las mujeres de todo el país, mientras que para los hombres sólo tiene importancia en el NOA. También se observa que el lugar de nacimiento es muy importante en la Patagonia y algo menos en GBA, independientemente del género de la persona. La mayor relevancia del régimen de tenencia de la propiedad parece estar en la región pampeana, mientras que el lugar de residencia cinco años atrás parece afectar en mayor medida a los hombres, aunque no parece ser una variable extremadamente relevante. El resto de las variables parece tener una relevancia menor, tanto en hombres como en mujeres.

Todos estos resultados, que merecen un mayor análisis, sirven para ilustrar la utilidad de los árboles de regresión para interpretar la información que surge de los datos. Más aún, ejemplifican el potencial de los mismos para identificar interacciones entre las variables que no habían sido consideradas a priori. Si bien para los objetivos de este trabajo lo más relevante es que en ningún caso se encontró un árbol vacío, lo que significa que siempre se rechazó la hipótesis nula de existencia de igualdad de oportunidades, el hecho de que sirvan para realizar análisis más complejos ayuda a justificar la utilización de este tipo de metodología.

En la próxima subsección se presentan los resultados de los bosques aleatorios, los cuales producen predicciones más precisas, las cuales serán utilizadas para estimar el nivel de desigualdad de oportunidades y el grado de importancia de cada variable.

6.3.2. Nivel de desigualdad de oportunidades e importancia de las variables

Con la misma base de datos de la subsección anterior, se procedió a estimar la distribución de ingresos contrafactual mediante bosques aleatorios condicionales. Para cada trimestre y región, se estimó un árbol conformado por el ensamble de 500 árboles, limitando a 3 la cantidad de variables a considerar en cada nodo y trabajando con un nivel de significatividad del 1%.

Como fue explicado oportunamente, la fortaleza de los bosques aleatorios radica en que producen estimaciones insesgadas reduciendo considerablemente la varianza de los árboles individuales. De esta forma, se obtienen predicciones más confiables y se puede cuantificar la importancia de cada variable de una manera mucho más precisa que la que se puede hacer observando los árboles. Para cada una de las 60 estimaciones se calculó la estimación de las variables mediante las técnicas explicadas en el Capítulo 4 y los resultados se normalizaron para que la variable más relevante tenga un valor de 100. Lo que se presenta en el Gráfico 30 es el promedio a lo largo de los 10 trimestres de la importancia de cada variable para cada región. En el anexo online está disponible el cálculo de la importancia de las variables para los diez trimestres de las seis regiones.

Los resultados están en línea con los de la sección anterior. En 4 de las 6 regiones la variable más relevante es el género de la persona, seguida por el tamaño del hogar. En el NOA estas dos variables también son las más relevantes, aunque en orden inverso. Por su parte, en la Patagonia el género es seguido en importancia por el lugar de nacimiento. Tal como se apreciaba en la sección anterior, el lugar de nacimiento también tiene una importancia relativa en el GBA. Las otras variables tienen una importancia menor, la cual varía de una región a otra.

Una vez estimada la distribución del ingreso contrafactual, aplicando el coeficiente Gini sobre ella se puede obtener un indicador del nivel de desigualdad de oportunidades para cada región. En la Tabla 10 y Gráfico 31 se presentan estos resultados.

Gráfico 30. Importancia de las variables promedio en los 10 trimestres

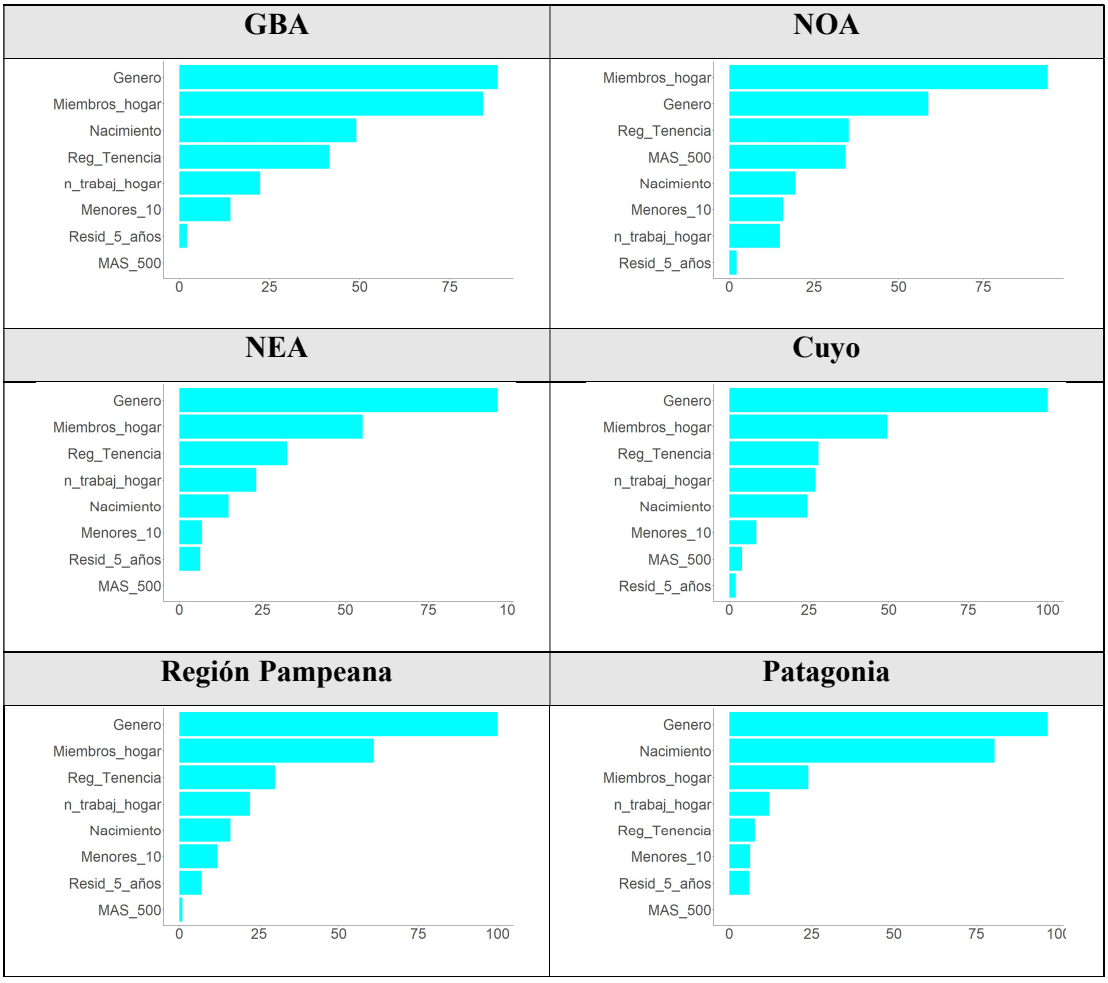


Tabla 10. Resumen de la desigualdad de oportunidades estimadas para región

Región	Desig. De Oport. Promedio	Desvío Standard
Cuyo	0,137	0,0137
GBA	0,159	0,0101
NEA	0,132	0,0113
NOA	0,138	0,0086
Pampeana	0,147	0,0048
Patagonia	0,136	0,0093

Puede observarse que en el GBA se registran mayores niveles de desigualdad de oportunidades que en el resto del país, ocupando el segundo lugar la región pampeana. Si bien los resultados no son comparables ya que se trabajó con distintas variables, recordemos que los niveles de desigualdad estimados por Brunori et al (2018) con la misma metodología para Suecia y Alemania eran de 0,03 y 0,08 respectivamente. Para todas las regiones del

país, acá se están estimando mayores niveles de desigualdad de oportunidades que para los países europeos.

Gráfico 31. Nivel de desigualdad de oportunidades por región

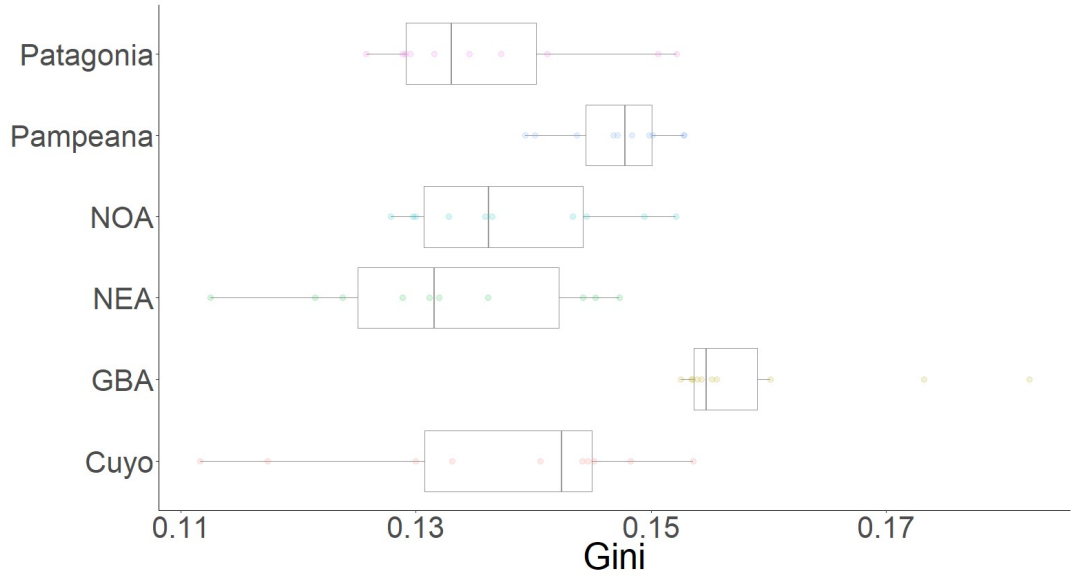
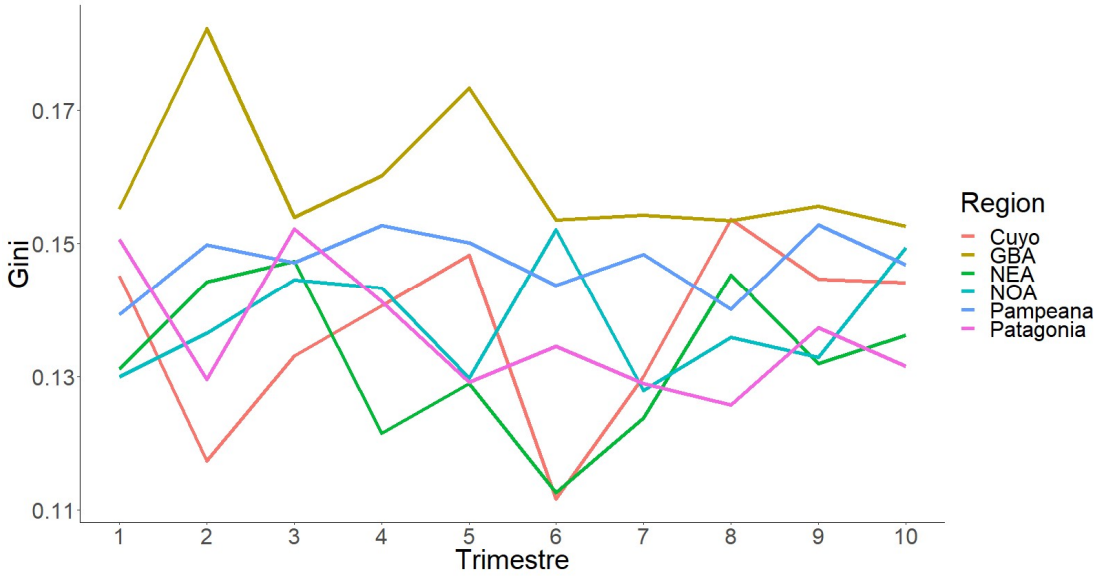


Gráfico 32. Evolución de la desigualdad de oportunidades en los 10 trimestres



El Gráfico 32 muestra que, a lo largo de los diez trimestres, las mediciones muestran cierta volatilidad. Sólo en la región pampeana los resultados son relativamente estables en torno a valores de 0,14-0,15. En GBA, por su parte, luego de unos trimestres volátiles el valor se estabiliza levemente por encima del 0,15. Para las demás regiones es difícil sacar

conclusiones concretas dada la volatilidad de los resultados. Sólo puede mencionarse que, en general, presentan menores niveles de desigualdad de oportunidades que en el GBA y la región Pampeana.

6.4. Resumen de los resultados obtenidos

Los resultados que se presentaron en este capítulo están en línea con lo esperado y con los objetivos e hipótesis de este trabajo. En la primera sección se mostró evidencia de que mediante árboles de regresión y bosques aleatorios se pueden obtener estimaciones del nivel de ingresos más precisas que las de otras metodologías tradicionales. La superioridad de estos métodos de ML es más evidente cuando la cantidad de variables y valores posibles de las mismas es muy grande en comparación a las observaciones disponibles, aunque también mostraron resultados más precisos para valores más grandes del ratio N/p. De esta forma, la metodología quedó validada para realizar las estimaciones de las secciones segunda y tercera.

En todos los casos analizados en las secciones 6.2 y 6.3 se rechazó la hipótesis nula de existencia de igualdad de oportunidades. En la mayoría de los casos se estuvo trabajando con un nivel de significatividad del 1%, con la salvedad de la sección 6.2.2, donde el tamaño reducido de la muestra obligó a reducir el nivel de significatividad al 10%. Por lo tanto, se puede afirmar que existe evidencia de la existencia de desigualdad de oportunidades en la Argentina y al interior de sus regiones.

Los resultados de la segunda sección muestran que el nivel educativo de los padres condiciona el nivel de ingreso de las personas, lo que es un claro signo de desigualdad de oportunidades y representa un límite a la movilidad social. Por su parte, en la tercer sección se muestra evidencia clara de que el género es otro factor determinante a la hora de explicar la distribución de los ingresos. Más aún, si bien el tamaño del hogar también surge como una variable sumamente relevante, parece afectar en mayor medida a las mujeres. Finalmente, en algunas regiones del país el lugar de nacimiento también es relevante para explicar la distribución de los ingresos. Mientras que en GBA los nativos parecen contar con ventaja, en el resto del país y en particular en la Patagonia, los nativos pierden respecto a los que provienen de otras provincias. Es de destacar que estas interacciones fueron sugeridas por el algoritmo empleado en base a los datos, no habiendo sido planteadas a la hora de formular el modelo. Esto sirve para ejemplificar la utilidad de estos métodos para hallar interacciones entre las variables que, a priori, habían sido ignorados.

Las estimaciones del nivel de desigualdad de oportunidades sugieren que el GBA es la región del país con mayor desigualdad, seguida por la región pampeana. De todas formas, los valores calculados tienen alta volatilidad, por lo que sería interesante probar la metodología en un período de tiempo más largo o con una base de datos que incluya una mayor cantidad de variables intergeneracionales. En este sentido, este trabajo constituye un punto de partida para futuros trabajos de mayor complejidad.

7. Conclusiones y reflexiones finales

El objetivo general de este trabajo era "estudiar si existe desigualdad de oportunidades en la Argentina, identificando sus principales determinantes". Cumplir con este objetivo requirió varios pasos, tanto a nivel teórico como empírico. En primer lugar, fue necesario adoptar una definición de igualdad de oportunidades entre las distintas que se encuentran en la literatura. En particular, se adoptó la definición *ex – ante*, según la cual existe igualdad de oportunidades cuando la distribución del ingreso es independiente de las circunstancias. En segundo lugar, fue necesario seleccionar cuáles serían las circunstancias a considerar. Esta selección estuvo principalmente basada en los datos disponibles, utilizando al menos tres variables que, sin discusión alguna, se encuentran fuera del control individual: género, lugar de nacimiento y, en la medida que la disponibilidad de datos lo permitió, nivel educativo de los padres.

Una vez adoptada una definición de igualdad de oportunidades y seleccionadas las variables a considerar, se planteó la discusión sobre cuál es la metodología más apropiada para realizar el análisis. En particular, uno de los objetivos específicos de este trabajo era analizar si mediante técnicas de *machine learning* se conseguía mejorar la calidad de las estimaciones de la distribución del ingreso. En base a los resultados obtenidos, se puede concluir que sí, al menos cuando se quiere considerar un gran número de variables en relación a la cantidad de observaciones. Dentro de las distintas alternativas que ofrecen estos métodos estadísticos/computacionales, se encontró que los árboles y bosques condicionales presentaban un marco conceptual ideal para testear la existencia de igualdad de oportunidades, dado que están basados en tests de hipótesis.

Mediante los distintos análisis realizados se presentó evidencia de que, en Argentina, no existe la igualdad de oportunidades ni a nivel país ni regional. Se consiguió identificar a los principales determinantes de la desigualdad de oportunidades: el nivel educativo de los padres, el género, el tamaño del hogar y, en menor medida, el lugar de nacimiento. Respecto

a esta última variable, se encontraron interesantes diferencias a nivel regional, lo cual posiblemente esté vinculado a los distintos procesos demográficos y económicos en curso en cada región. También se encontró una importante interacción entre el género de una persona y el tamaño del hogar. Pareciera ser que las mujeres sufren dos tipos de desventajas a la hora de obtener ingresos: ser mujer y tener que soportar las cargas de vivir en un hogar con muchos integrantes.

Respecto a estas interacciones, tal vez el punto a resaltar no sea la interacción en sí, dado que suenan lógicas y seguramente no sorprenden a alguien que sigue de cerca las cuestiones distributivas y regionales de nuestro país. Lo que considero relevante de estas interacciones es la forma en que fueron halladas. No fueron planteadas previamente por el autor de este trabajo, sino que fueron encontradas por el algoritmo en base a los datos. Esto ejemplifica el gran aporte que pueden hacer las metodologías empleadas en este trabajo: identificar patrones o interacciones que, a priori, habían sido ignorados.

El último ejercicio realizado consistió en estimar el nivel de desigualdad de oportunidades mediante el coeficiente Gini tanto para Argentina como para cada una de las regiones. Estos resultados deberían ser tomadas como un ejemplo de cómo puede realizarse este tipo de estimación, aunque sería recomendable hacerlo con una base de datos que permita considerar una mayor cantidad de circunstancias. En particular, sería deseable incluir mayor cantidad de información referida al entorno familiar y social en el que nació y se desarrolló el individuo. Por ejemplo, sería interesante poder utilizar la base de un Censo Nacional de Población y Vivienda. Mejor aún, sería deseable que en los cuestionarios de la EPH se incorporen preguntas sobre los padres del jefe de hogar y su cónyuge, tales como el nivel educativo alcanzado. Esto permitiría un análisis más preciso sobre el nivel de desigualdad de oportunidades, realizando un seguimiento de la misma con frecuencia trimestral.

En base a todo lo anterior, considero que se han cumplido con los objetivos del trabajo, presentándose evidencia para sostener las hipótesis formuladas. Más aún, a lo largo del trabajo fue necesario abordar algunas cuestiones que, en principio, no estaban dentro de los objetivos del trabajo pero que enriquecen al mismo. Por ejemplo, la evidencia de que el desempeño relativo de los métodos basados en árboles mejora cuando la cantidad de variables es grande en relación a la cantidad de observaciones también puede considerarse un aporte relevante de este trabajo. Del mismo modo, el análisis realizado en la sección 6.2.1 respecto a cómo el árbol trata las diferencias regionales contribuye a entender las diferencias entre este tipo de algoritmos y las regresiones usualmente utilizadas en la literatura

económica, posibilitando una mejor comprensión sobre cuándo es conveniente utilizar una metodología o la otra.

En conclusión, considero que este trabajo contribuye a lograr una mayor comprensión respecto a cómo los métodos de *machine learning* pueden contribuir a la disciplina económica. Se presentó un análisis que permite una mejor identificación de los factores relevantes a la hora de analizar la problemática distributiva de nuestro país. Todo el trabajo fue realizado en base a datos de publicación periódica y de fácil acceso, sirviendo como punto de partida para nuevos estudios con mayor grado de profundidad.

8. Referencias bibliográficas

Alesina, A., & La Ferrara, E. (2005). Preferences for redistribution in the land of opportunities. Journal of public Economics, 89(5-6), 897-931.

Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Banco Mundial (2006). World Development Report 2006: Equity and Development. TheWorld Bank and Oxford University Press, Washington, DC (2006)

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, *36*(1-2), 105-139.

Bourguignon, F., Ferreira, F. H., & Menéndez, M. (2007). Inequality of opportunity in Brazil. *Review of income and Wealth*, *53*(4), 585-618.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. Chapman & Hall/CRC.

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

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics*, *26*(3), 801-849.

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

Brieuc, M. S., Waters, C. D., Drinan, D. P., & Naish, K. A. (2018). A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Molecular ecology resources*, *18*(4), 755-766.

Brunori, P., Peragine, V., & Serlenga, L. (2016). Upward and downward bias when measuring inequality of opportunity. *Social Choice and Welfare*, 1-27.

Brunori, P., Hufe, P., & Mahler, D. G. (2018). *The roots of inequality: Estimating inequality of opportunity from regression trees*. The World Bank.

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, *30*(4), 927-961.

Carrera, J.E., Rodríguez, E. (2013). Impactos macroeconómicos de la desigualdad. Congreso anual de la Asociación de Economía para el Desarrollo de la Argentina (AEDA). Buenos Aires, Septiembre de 2013.

Carrera, J. E., Rodriguez, E., & Sardi, M. (2016). The Impact of Income Distribution on the Current Account. Journal of Globalization and Development, 7(2).

Checchi, D., & Peragine, V. (2010). Inequality of opportunity in Italy. *The Journal of Economic Inequality*, *8*(4), 429-450.

Checchi, D., Peragine, V., & Serlenga, L. (2015, July). Income Inequality and Opportunity Inequality in Europe: Recent Trends and Explaining Factors. In *5th ECINEQ meeting, University of Luxembourg*.

Corak, M. (2012), Inequality from generation to generation: the United States in Comparison. Graduate School of Public and International Affairs, University of Ottawa.

DeBarr, D., & Wechsler, H. (2009, July). Spam detection using clustering, random forests, and active learning. In *Sixth Conference on Email and Anti-Spam. Mountain View, California*(pp. 1-6).

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, *40*(2), 139-157.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference*(Vol. 5). Cambridge University Press.

Faur, E. (2008). Desafíos para la igualdad de género en la Argentina. Estrategia del Programa de las Naciones Unidas para el Desarrollo. Buenos Aires: PNUD.

Ferreira, F. H., & Gignoux, J. (2011). *The measurement of inequality of opportunity: Theory and an application to Latin America*. The World Bank.

Friedman, M. y Friedman, R. (1979), *Libertad de Elegir.* Ediciones Orbis S.A., Madrid, 1983.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics. Second Edition (2008).

Gasparini, L. C. (2002). On the measurement of unfairness An application to high school attendance in Argentina. *Social Choice and Welfare*, *19*(4), 795-810.

Gholamian, E. y Davoodi, S. M. R. (2018). Predicting the direction of stock market prices using random forest.

Goldstein, B. A., Polley, E. C., & Briggs, F. B. (2011). Random forests for genetic association studies. *Statistical applications in genetics and molecular biology*, *10*(1).

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651-674.

INDEC (2003). La nueva Encuesta Permanente de Hogares de Argentina. 2003.

INDEC (2016). Consideraciones sobre la revisión, evaluación y recuperación de la Encuesta Permanente de Hogares (EPH). Anexo Informe de prensa, 23 de Agosto de 2016.

INDEC (2018). Valorización mensual de la canasta básica alimentaria y de la canasta básica total. Gran Buenos Aires. Informes Técnicos Vol. 2 n°137, Junio de 2018.

INDEC (2019). Encuesta Permanente de Hogares. Diseño de Registro y Estructura para las bases preliminares Hogar y Personas. Febrero de 2019.

Jiménez, M. (2016). *Movilidad intergeneracional del ingreso en Argentina: Un análisis de sus cambios temporales desde el enfoque de igualdad de oportunidades,* Documento de Trabajo, No. 203, Universidad Nacional de La Plata, Centro de Estudios Distributivos, Laborales y Sociales (CEDLAS), La Plata

Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*.

Lantz, B. (2013). *Machine learning with R*. Packt Publishing Ltd..Second Edition (2015)

Lefranc, A., Pistolesi, N., & Trannoy, A. (2009). Equality of opportunity and luck: Definitions and testable conditions, with an application to income in France. *Journal of Public Economics*, *93*(11-12), 1189-1207.

Liu, C., Chan, Y., Alam Kazmi, S. H., & Fu, H. (2015). Financial fraud detection model: based on random forest. *International journal of economics and finance*, *7*(7).

Marrero, G. A., & Rodríguez, J. G. (2009). Inequality of opportunity and growth.

Maurizio, R. (2008). Migración y desarrollo: el caso de Argentina. Migraciones internacionales en América Latina. Booms, crisis y desarrollo, 75184.

McCauliff, S. D., Jenkins, J. M., Catanzarite, J., Burke, C. J., Coughlin, J. L., Twicken, J. D., ... & Cote, M. (2015). Automatic classification of Kepler planetary transit candidates. *The Astrophysical Journal*, *806*(1), 6.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

Paes de Barros R., Ferreira F., Molinas Vega J., Saavedra Chanduvi J. (2009). Measuring Inequality of Opportunities in Latin America and the Caribbean. Banco Mundial.

Pérez, P. E. (2008). Desigualdades de género en mercado de trabajo argentino (1955-2003). Trabajos y comunicaciones.

Rawls, J. (1971). *A theory of justice*. Harvard university press. Edición de 2009.

Ramos, X., & Van de Gaer, D. (2012). Empirical approaches to inequality of opportunity: Principles, measures, and evidence.

Roemer, J. E. (1998). *Equality of opportunity* (No. 331.2/R62e). Cambridge, MA: Harvard University Press.

Sen, A. (1985). Commodities and capabilities. Amsterdam: North-Holland.

Serio, Monserrat (2011) : Igualdad de oportunidades en ingresos: el caso de Argentina, Documento de Trabajo, No. 126, Universidad Nacional de La Plata, Centro de Estudios Distributivos, Laborales y Sociales (CEDLAS), La Plata

Sohnesen, T. P., & Stender, N. (2017). Is Random Forest a Superior Methodology for Predicting Poverty? An Empirical Assessment. *Poverty & Public Policy*, *9*(1), 118-133.

Sosa Escudero, W. (2019). Big data. Siglo XXI Editores, Buenos Aires.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, *14*(4), 323.

Vallentyne, P. (2002). Brute luck, option luck, and equality of initial opportunities. *Ethics*, *112*(3), 529-557.

Van De Gaer, D. F. G. (1995). Equality of opportunity and investment in human capital.

Zhou, L., & Wang, H. (2012). Loan default prediction on large imbalanced data using random forests. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, *10*(6), 1519-1525.

9. Anexo.

Toda la parte analítica de este trabajo se realizó mediante el lenguaje de programación "R"[29], el cual se distribuye mediante la licencia de código abierto "GNU GPL"[30].

Gran parte del código fuente utilizado en este trabajo, como también algunos resultados auxiliares, se encuentran disponible en el Anexo Online disponible en el siguiente repositorio: https://github.com/esterodr/DO

En particular, en el Anexo Online se presentan los siguientes archivos:

Link	Descripción
generar_bases.R	Descarga los archivos de la EPH de la página del INDEC y realiza las transformaciones necesarias para generar la base de datos utilizada.
seccion6_1_casos_1_y_2.R	Realiza las estimaciones de los Casos 1 y 2 incluidos en la sección 6.1.
seccion6_1_caso_3.R	Realiza las estimaciones del Caso 3 incluido en la sección 6.1.
seccion6_2.R	Realiza las estimaciones cuyos resultados se presentan en la sección 6.2.
seccion6_3.R	Realiza las estimaciones cuyos resultados se presentan en la sección 6.3.
Árboles de oportunidad – Trimestres 2 a 9.pdf	Resultados de los árboles de regresión condicionales no publicados en la sección 6.3.1 por falta de espacio.
Importancia de las Variables por trimestre y region.pdf	Importancia de las variables estimada por los bosques aleatorios condicionales por trimestre y región. El promedio de estos resultados es lo que se publica en la sección 6.3.2.

[29] https://www.r-project.org/
[30] https://www.gnu.org/licenses/old-licenses/gpl-2.0.html