# Sección 3
# El Aprendizaje Automático: un enfoque metodológico a los programas sociales

Guy Lacroix, Luis Huesca y Linda Llamas

Los métodos de Aprendizaje Automático (AA) son técnicas que parten de la inteligencia artificial y permiten cuantificar los efectos de las variables en estudio y la manera en que inciden de forma directa e indirecta en determinados programas de políticas públicas.

En este estudio lo utilizamos como un método innovador y alternativo a las técnicas de análisis tradicionales, cuyo fin es medir el impacto en el bienestar de las mujeres apoyadas por el Programa en su conjunto, así como en las diversas vertientes que lo componen. Esto permitirá medir su efecto en la generación de bienestar y conocer en qué medida las mujeres beneficiadas pueden elevar su calidad de vida a nivel individual.

Si bien los métodos de AA han recibido mucha atención en los últimos años, los mismos son utilizados, principalmente, para realizar predicciones. Investigadores empíricos que realizan evaluaciones de políticas públicas están preocupados por los problemas causales, tratando de responder preguntas contrafactuales, por ejemplo ¿qué hubiera pasado en ausencia de una política?

Debido a que en este tipo de análisis se dificulta observar directamente los resultados en el momento de su aplicación, no al menos hasta que suceda en la práctica, la herramienta de predicción del ML permite producir innovaciones, incorporando herramientas en estimadores de parámetros causales, como el ATE, explicado en Heckman, *et al.* (1998).

Afortunadamente, autores han ayudado recientemente a desarrollar métodos de ML para el análisis causal (Athey, 2017a, 2017b; Athey e Imbens, 2017; Kleinberg, Ludwig, Mullainathan y Obermeyer, 2015; Kreif y Diaz-Ordaz, 2019; Mullainathan y Spiess, 2017; Varian, 2014). Estos métodos, también conocidos como *Random Forests, LASSO, Ensemble method*, constituyen un campo de investigación activo en la literatura

científica, y son utilizados cada vez más por diferentes Secretarías de Estado, así como instituciones académicas privadas o públicas.

Lo anterior nos permitirá, o incluso, eliminar primero los problemas de especificación de modelos econométricos de tipo más tradicional que puedan llevar a resultados en ocasiones erróneos y aumentar la transparencia en la selección del modelo.

La nueva literatura incluye técnicas que incorporan enfoques del ML en la estimación del ATE de un tratamiento binario –explicado este último de forma amplia en Heckman, *et al.*, (1998)–, pero añadiendo el toque del criterio de la desconfianza y la positividad de un problema.

Además, ofrece una alternativa a las funciones no paramétricas de ponderación. Los métodos de ML se han aplicado con éxito para trazar la heterogeneidad del efecto del tratamiento en grupos donde se busca evaluar el impacto de un programa.

Esto permite identificar qué subpoblación se beneficia más del tratamiento, en nuestro caso, las mujeres que han sido apoyadas con respecto a su característica de grupo (si residen en zonas urbanas o rurales, número de hijos, etc.), algo que los métodos clásicos no pueden hacer, así como conocer de primera mano las expectativas de éxito en el mercado laboral o de salir de su condición de pobreza.

Es importante mencionar que el BM está interesado también en utilizar los métodos de ML para estudiar la pobreza, particularmente para identificar la heterogeneidad en los efectos del tratamiento de uno o varios programas de política pública, técnica que va un paso delante de la tradicional o estándar de obtención de los efectos tratamiento tipo ATT o ATE en econometría.

Para realizar el análisis comparativo usando el ML se considerarán estimaciones realizadas a partir de la revisión de variables de la ENOE, señaladas en la sección 2, para ahondar en la generación de evidencias y su interpretación en el colectivo femenino de interés. Con base en los hallazgos se generarán los indicadores para realizar las mediciones, el monitoreo y el seguimiento del bienestar, la calidad de vida y las brechas a partir del Salario Rosa, así como la inclusión de recomendaciones que faciliten mayores niveles de bienestar a nivel individual y colectivo.

Como se argumentó en la sección anterior, es importante advertir que la vertiente Salario Rosa por la Cultura Comunitaria es de carácter subjetivo y podría ser de difícil medición, en el entendido de que carece de datos tangibles identificables en la encuesta. La forma

más adecuada de medir las vertientes será la permitida por el proceso de los bosques aleatorios (*Random Forest*); a continuación, se explica esta técnica de medición.

## 3.1. Los Bosques Aleatorios Causales (BAC)

En esta sección, se muestra brevemente una explicación del concepto de los bosques aleatorios (*Random Forest* en ML) que sintetizan los resultados del AA. Mostramos cómo se pueden utilizar para estimar los efectos causales del tratamiento dentro del Programa de Desarrollo Social FFSR. El énfasis está en la intuición del método. Las propiedades detalladas y la rigurosa teoría asintótica, del cual dependen, están fuera del alcance de este apartado; sin embargo, sugerimos consultar los trabajos de Breiman (2001) y Athey & Wager (2019), los cuales abordaremos en los siguientes párrafos.

### *3.1.1. El marco e intuición de los BAC*

Breiman (2001) introdujo el desarrollo de los Bosques Aleatorios Causales (BAC); en sus inicios, el concepto proporcionó una forma no paramétrica de clasificar objetos e implementar modelos de regresión para la inferencia, al mismo tiempo ofrecía información gráfica intuitiva y fácil de interpretar.

Un BAC es una colección de árboles compuestos de nodos, ramas y hojas. Los árboles de regresión, que normalmente se utilizan para estimar modelos predictivos en el Aprendizaje Automático, nos permiten resaltar las características más importantes dentro de un modelo.

Funcionan bien con grandes conjuntos de datos y con un gran número de covariables. Una ventaja que tiene este método, por sobre las regresiones paramétricas de tipo *Propensity Score Matching (PSM)* y más tradicionalmente utilizadas en este análisis, es que el investigador no está obligado a decidir sobre los cortes arbitrarios para una variable continua.

Con los BAC, las opciones de cortes están basadas en los propios datos. Además, no supone *per se*, que la relación entre el tratamiento y las covariables empleadas sea lineal, como lo es con Heckman, *et al.* (1998), ni supone que la probabilidad de tratamiento siga una distribución normal (tipo probit) o logística (tipo logit).

La figura 4 presenta un árbol, el cual se construye de arriba hacia abajo: el algoritmo selecciona primero una variable (un primer nodo) que discrimina mejor entre los diferentes resultados en función de alguna medida de entropía clásica (como podría ser en la desigualdad, índices de Theil, o incluso el Gini).

En nuestro caso, el algoritmo implementa la división que minimiza un criterio de bondad de ajuste en la Σ de la muestra, utilizando el Error Cuadrado Medio (MSE) como $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$, donde $\hat{y}_i$ es la Y promedio dentro de la hoja de un árbol de observación.

A continuación, el algoritmo repite el proceso para cada una de las dos hojas nuevas, y así sucesivamente hasta que alcanza una regla de detención. Por lo tanto, al elegir una división, el algoritmo busca maximizar la diferencia en el efecto de tratamiento entre los dos nodos secundarios.

El nodo entonces genera dos bifurcaciones, una si se cumple la condición dada por el nodo y otra si no lo es. La primera minimiza la distancia al interior de un nodo, y la segunda maximiza la distancia entre dos nodos.

El algoritmo se aplica nuevamente para encontrar otros nodos. Las bifurcaciones se detienen cuando se logra un nivel máximo de profundidad o debido a que el número de observaciones en las hojas (los nodos terminales) alcanzan un valor preespecificado. La profundidad a la que nos referimos es un parámetro que establecemos, así como el número de observaciones por hoja que también es un parámetro que fijamos.

Como se mencionó, los bosques son una agregación de muchos árboles y se emplean para reducir la varianza del estimador. Athey (2019) desarrolló las nuevas técnicas que permiten a los investigadores extraer inferencias causales de bosques aleatorios que se utilizaron inicialmente para predecir (no explicar) los resultados. Describimos brevemente algunas características del método a continuación.

### 3.1.2. La agregación de los BAC con Bootstrap

La agregación de BAC con *Bootstrap* permite mejorar la clasificación mediante la incorporación de varios árboles. En cada iteración se utiliza una muestra diferente de los datos originales para hacer crecer el árbol. Además, el algoritmo utiliza la selección de división aleatoria: en cada nodo se utiliza un subconjunto diferente de variables exógenas que

llamaremos $m$, con la condición $m < K$, para dividir el nodo en dos ramas. Este paso produce árboles menos correlacionados y, por tanto, reduce la varianza[7].

#### 3.1.2.1. La honestidad en la formación del árbol y su impacto

Wager y Athey (2018) introdujeron el concepto de bosques aleatorios honestos. Éste cultiva árboles, usa la mitad de la muestra de arranque y la otra se utiliza para estimar los efectos de tratamiento de interés. En nuestro cálculo consideramos la honestidad sugerida por los autores.

#### 3.1.2.2. La Partición del BAC

Los bosques aleatorios causales son adecuados para explorar la heterogeneidad que, sobre todo, se dificulta captar en cualquier tipo de programa o política pública de interés. El proceso de división está optimizado para capturar efectos de tratamiento heterogéneos. Athey (2019) muestra cómo explotar la partición recursiva mediante la definición de un nuevo criterio que aumenta la heterogeneidad en los efectos del tratamiento lo más rápido posible con el uso de esta técnica.

### 3.1.3. El BAC en el Programa de Desarrollo Social FFSR: un ejemplo de árbol sintetizado

Este árbol se hizo con una muestra aleatoria extraída de todos los datos en el Programa, en función de identificarlos en la base de datos de la ENOE, correspondiente al primer trimestre de 2018. Según el algoritmo, tener cuatro hijos es determinante en el Programa y aparece como la variable discriminante más importante, responsable de inducir un efecto de tratamiento en el Programa de Desarrollo Social FFSR y sus componentes no monetarios.

A continuación, el algoritmo identifica seguidamente a la escolaridad (años formales educativos), como la principal variable discriminatoria. Así, las mujeres con dos años de escolaridad conforman el mayor impacto (incluido en la hoja violeta del árbol). La hoja contiene la cifra de mujeres similares en términos de hijos y escolaridad, se muestra que 55.0% de ellas son pobres y 45.0% se benefician del Salario Rosa.

[7] Véase Ishwaran (2015) para más detalles acerca de la técnica.

Dichos datos (que son válidos solo para este árbol en cuestión) son obtenidos por el proceso matemático que replica tantas veces como sea necesario para llegar a ellas, y son dados precisamente por la estructura de la información y sin tener que realizar cálculos adicionales acerca de la pobreza y/o de la cobertura. Es probable que la primera división para otro árbol sea distinta.

El impacto estimado del Programa de Desarrollo Social FFSR para la hoja en cuestión se calcula como la diferencia entre la proporción de mujeres pobres para los grupos tratados y de control –de manera análoga a como se hace con el *propensity score*– (Heckman, *et al.*, 1998).

La contribución de la hoja a la estimación global se pondera por el número de observaciones que contiene. Las mujeres que tienen años de escolaridad distintos a dos, se dividen en la próxima división entre las que viven en un área urbana y las que no.

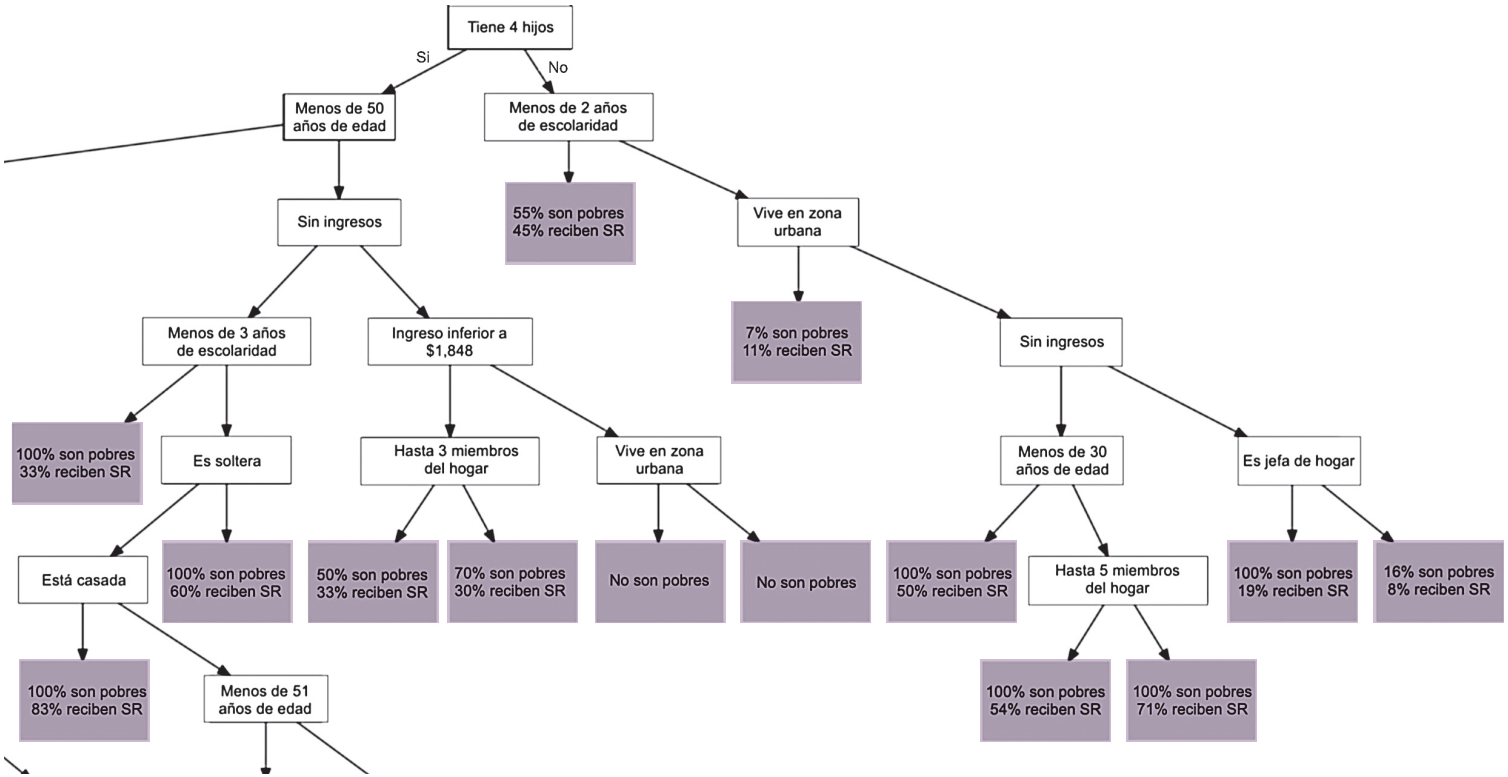
Estos últimos conforman la hoja violeta, y en ella la proporción de mujeres pobres es de 7.0%, mientras que la proporción de beneficiarias del Programa de Desarrollo Social FFSR en este subconjunto es de 11.0%; es decir, el Programa capta solo este porcentaje con respecto a la atención en ese 7.0% de mujeres en condición de pobreza.

A medida que bajamos del árbol, nos encontramos con una serie de hojas en las que la tasa de pobreza va de 0.0% a 100.0%, independientemente de la proporción de mujeres que reciben Salario Rosa (o alguna de sus vertientes).

Estas hojas aportarán un efecto de tratamiento de 0 a la estimación global, cuyo peso dependerá del número de observaciones que contengan cada una.

Los árboles pueden contener muchos nodos y hojas. A menudo, son simplemente demasiado grandes para ser representados gráficamente y en la medida que se estima un árbol aleatorio, una mujer en tratamiento caerá en diferentes hojas. El efecto de tratamiento para esa persona será entonces el resultado de un promedio calculado sobre todos los árboles contenidos en el bosque.

## Figura 4. Bosques Aleatorios Causales (BAC) e interpretación




Fuente: elaboración propia con base en simulación, empleando el Programa R y base de datos de proyecto.