Efectos Heterogéneos en Sindicalización, un Enfoque desde Machine Learning.

Leonardo Lanata

Borrador final de tesis presentado para el seminario de
Magíster en Análisis Económico.

Departamento de Economía
Universidad de Chile
Chile
9 de Diciembre, 2022

Efectos Heterogéneos en Sindicalización, un Enfoque desde Machine Learning

Leonardo Lanata[1,2]

[1] Facultad de Economía y Negocios, Universidad de Chile
leonardolanata@gmail.com
[2] Supervisor: Esteban Puentes

Bullet Points

1. Usamos *Machine Learning* para estimar efectos heterogéneos en los salarios por sindizalicación.
2. En este contexto, mostramos las ventajas de estas metodologías por sobre la econometría tradicional.
3. Ocupando *Meta-Learners*, pudimos comprobar heterogeneidad y supuestos clave de balance.
4. Encontramos un ATE sobre salarios igual al 3.6%.
5. Computamos un modelo interpretativo de los algoritmos, usando valores SHAP.

Abstract. We study heterogeneity in the effect of union on wages through an observational dataset from the Encuesta de Protección Social (EPS) in Chile. The analysis uses Machine Learning to address the following associated problems: assessing treatment group positivity and covariate balance, imputing conditional average treatment effects, and interpreting imputed models. By comparing several different *meta-learners* models we illustrate the flexibility of these estimators in regards to predict treatments effects. We find that unionization of workers has a possitive effect on wages of 3.6%, 95%-CI of [3.02%,4.18%], and that the heterogeneity in the range of [-21%,65%] is determined by industries, firm size, experience, occupations, and gender.

Abstract. Estudiamos la heterogeneidad en el efecto de la sindicalización sobre los salarios a través de un conjunto de datos observacionales de la Encuesta de Protección Social (EPS) en Chile. El análisis utiliza *Machine Learning* para abordar los siguientes problemas: evaluar el supuesto de positividad de los grupos de tratamiento y el balance de covariables; computar efectos condicionales promedio del tratamiento e interpretar los modelos imputados. Al comparar varios modelos diferentes de *meta-learners*, ilustramos la flexibilidad de estos estimadores en lo que respecta a predecir los efectos de los tratamientos. Encontramos que la sindicalización de los trabajadores tiene un efecto positivo sobre los salarios de 3.6%, con un intervalo de confianza al 95% de [3.02%,4.18%], y

que la heterogeneidad contenida en el rango de [-21%,65%] está determinada por industrias, tamaño de la firma, experiencia laboral, ocupaciones y género.

Keywords: Machine Learning · Interpretability · Counterfactual estimation.

1 Introducción y motivación

El *Machine Learning* ha tenido un éxito ampliamente aceptado en la resolución de problemas de predicción. En particular, hay aplicaciones que van desde el reconocimiento de imagen y voz (LeCun et al., 2015) [27] hasta la predicción de enfermedades en medicina (Kononenko, 2001) [11]. En particular el aprendizaje supervisado va en torno a resolver algún problema de predicción, en el sentido de que produce un predicho $\hat{y}$ a partir de un set de covariables (o *features*) x. En general, la economía aplicada va más de buscar relaciones causales entre variables, por lo que esta premisa asume encontrar un problema $\hat{y}$ relevante si es que quisiéramos añadir *Machine Learning* al conjunto de herramientas que la econometría proporciona para resolverlos. El atractivo de estas metodologías viene del hecho de poder descubrir patrones generalizables, de estructura compleja y en formas funcionales flexibles de los datos. Todo esto, sin caer en el problema del *overfitting*: estas herramientas pueden encontrar funciones que predicen apropiadamente fuera de la muestra.

En la literatura podemos revisar varias aplicaciones de *Machine Learning* en economía. En particular, estos modelos pueden ser aplicados resolviendo preguntas tradicionales pero con nuevos conjuntos de datos; por ejemplo, podemos medir la actividad económica utilizando imágenes satelitales a través del tratamiento de imágenes o en base a titulares de las noticias con técnicas de procesamiento de lenguaje natural [3]. En el caso de Chile, son muy pocas las aplicaciones documentadas. En Contreras et al., 2018 [6] estimaron varios modelos de *Machine Learning* para predecir el medio por el cual los individuos se movilizan hacía su trabajo y con sus predicciones generaron *proxies* para el tiempo de viaje de los trabajadores. En Kausel et al., 2018 [10] utilizaron métodos de aprendizaje para determinar que la relación ancho a alto de la cara podría predecir el desempeño académico de los estudiantes de economía y negocios, donde encontraron que rasgos dominantes en varones podrían tener un efecto positivo en el desempeño de asignaturas de carácter cualitativo. De manera más reciente, Leal et al., 2020 [12] estimaron la inflación para Chile empleando variados métodos de *Machine Learning*, donde encontraron que en términos predictivos estos modelos se desempeñan de manera similar a los modelos internos del banco, sin embargo en términos estructurales estas metodologías no tenían mucho que aportar de-

[3] Para ver varias aplicaciones en economía internacional, véase Mullainathan y Spiess, 2017 [15]

bido a su dificultad interpretativa.

En este trabajo nos enfocaremos en aplicar estas herramientas para estudiar la heterogeneidad en los efectos causales de la sindicalización de los trabajadores. De hecho, en términos teóricos, las metodologías de *Machine Learning* sobresalen en la superación de las conocidas limitaciones de los métodos tradicionales que emplea la econometría para resolver este tipo de problemas. Por ejemplo. los métodos de *matching* no suelen tener un buen desempeño cuando el set de covariables es de varias dimensiones (Rubin y Thomas, 1996) [7]; los modelos lineales generalizados no son lo suficientemente flexibles para descubrir interacciones variables y tendencias no lineales; y los métodos basados en *propensity* tienen problemas de alta-varianza en la estimación (Lee et al., 2011) [5]. Por el contrario, el aprendizaje supervisado ha demostrado ser útil para descubrir patrones en datos de alta dimensionalidad (Lecun et al., 2015) [27], aproximar funciones complejas y en el intercambio (*trade-off*) entre sesgo y varianza (Swaminathan y Joachims, 2015) [1].

En este estudio observacional basado en datos de la Encuesta de Protección Social de Chile, aplicamos estimadores de dos etapas de la familia de modelos *Meta-Learners*, diseñados específicamente para caracterizar heterogeneidad en el efecto de la sindicalización sobre el salario de los trabajadores. La heterogeneidad final se computará empleando el modelo de mejor desempeño. Considerando esto, este estudio es relevante y aporta a la literatura en el sentido de aplicar y caracterizar un nuevo marco metodológico para la evaluación empírica de programas en Chile, utilizando *Machine Learning*. En las siguientes secciones, describiremos nuestro problema, revisaremos la metodología de estimación y los resultados, finalmente, interpretaremos el modelo.

2 Metodología y Datos

2.1 Descripción del problema y los datos

Estudiamos el efecto de la sindicalización en los salarios de los trabajadores, basados en datos de la Encuesta de Protección Social (EPS) de Chile. Para la edición de 2002-2009, la muestra comprende un total de 3462 observaciones. Esta base de datos cuenta con información a nivel individuo, de su salario, experiencia laboral, industria y tamaño de la firma en la que trabaja, si es que este tiene algún tipo de limitación física y entre otras. La sindicalización (desde ahora el tratamiento) es representada por una variable binaria $T \in \{0, 1\}$, mientras el *outcome*, $Y \in R$ representa al logaritmo del salario por hora. Los trabajadores son observados a través de las covariables X_0 que corresponde a la experiencia en años; X_1 que representa categóricamente el nivel educacional, siendo el primero nivel menor a 12 años, luego exactamente 12 y para más de 12 años; X_2, X_3, X_4 y X_5 son variables dicotómicas que identifican respectivamente a trabajadores de mediana habilidad, si tienen alguna restricción física, su género y si son trabajadores del sector privado o no; X_6 corresponde al tamaño de la empresa,

comenzando el primer tramo entre 10 y 49 trabajadores, luego entre 50 y 199, para finalmente clasificar a las firmas mas grandes con más de 500; X_7 describe las posibles ocupaciones (profesionales, técnicos, administrativos, venta y servicios, etc); por último X_8 describe la industria del trabajador, ya sea manufactura, minería, construcción, etc. Por conveniencia, dejamos a $X = [X_1, X_2, ..., X_8]^T$ representar el set completo de covariables observables para los trabajadores de la muestra. Así, dejamos a $\{x_i, t_i, y_i\}$ denotar la observación correspondiente para el individuo $i \in \{1, ..., n\}$. Los grupos del tratamiento observados los denotaremos como G_0 para el control y G_1 para los tratados y se definen por $G_z = \{i \in \{1, ..., n\} : t_i = t\}$. El data set completo de observaciones lo denotamos $\mathcal{D} = \{(x_1, z_1, y_1), ..., (x_n, t_n, y_n)\}$ y la densidad de las variables como $p(X, T, Y)$.

Con el problema descrito, podemos adoptar el enfoque de modelos causales de Neyman-Rubin (Rubin, 2005) [19], y denotar por $Y(0), Y(1)$ los *outcomes* potenciales correspondientes a la intervención $T = 0$ y $T = 1$ respectivamente. El objetivo de este estudio es caracterizar la heterogeneidad en el efecto del tratamiento $Y(1) - Y(0)$ a través de las variables observables X que recién definimos para los trabajadores, las cuales corresponden a posibles fuentes de heterogeneidad en el tratamiento (ver por ejemplo (Landerretche et al., 2013) [17]). Como es bien sabido, este efecto no es identificable sin supuestos adicionales dado que cada trabajador es observado solamente en uno de los grupos de tratamiento. En cambio, estimaremos el *efecto condicional promedio del tratamiento* (*CATE* por sus siglas en inglés) con respecto a las covariables observadas X.

$$\tau(x) = E[Y(1) - Y(0) \mid X = x] \tag{1}$$

La ecuación 1 es identificable a partir de datos observacionales bajo el supuesto estándar de ignorabilidad,

$$Y(1), Y(0) \perp T \mid X,$$

consistencia, $Y = ZY(1) + (1 - Z)Y(0)$ y positividad,

$$\forall x : p(Z = 0 \mid X = x) > 0 \iff p(Z = 1 \mid X = x) > 0$$

El CATE condicionado al set completo de covariables X es lo más cerca que estaremos de estimar el efecto del tratamiento para un trabajador individual. Sin embargo, para el diseño y asignación de políticas, es poco frecuente -y necesario- tener este nivel de resolución. En secciones más avanzadas de este trabajo, siguiendo a Johansson (2018) [9] estimaremos los efectos condicionales también con respecto a las variables individuales de X, para obtener por ejemplo, el efecto promedio marginal estratificado por industria, ocupación, o cualquier sub-muestra o función de X.

2.2 Marco Teórico y Metodológico

En esta sección primero expondremos a grandes rasgos el marco unificado de meta-algoritmos (meta-learners) propuesto en Künzel et al., 2019 [23] que com-

prende el estado de arte en cuanto a estimación de heterogeneidad en efectos de un tratamiento, a través de aprendizaje supervisado. Luego revisaremos el *pipeline* estándar de *machine learning* mediante el cual describiremos el proceso de división de la base de datos en la sub-muestra de entrenamiento y validación, computaremos la estimación de los *outcomes* potenciales y caracterizaremos la heterogeneidad en los CATE.

Meta-Learners

Existe un creciente interés por estimar y analizar efectos heterogéneos de un tratamiento en estudios tanto experimental y observacionales. Siguiendo a Künzel et al., 2019 [23] describiremos algunos meta-algoritmos que pueden sacarle ventaja a cualquier método de aprendizaje supervisado o de regresión para estimar los CATEs en un contexto de asignación binaria del tratamiento. La idea central de los meta-algoritmos es descomponer la estimación de los CATEs en varios sub-problemas de regresión que pueden ser resueltos con cualquier regresión o método de aprendizaje supervisado.

El meta-algoritmo más común para estimar la heterogeneidad en el efecto del tratamiento comprende dos pasos. Primero, emplea un algoritmo de aprendizaje de base (por ejemplo, regresiones lineales o métodos basados en árboles de decisión) para estimar los *outcomes* condicionales para individuos bajo control y aquellos bajo tratamiento. Segundo, en la muestra de validación, toma la diferencia entre estas estimaciones para computar los CATEs. En general, nos referimos a este mecanismo de estimar por separado como *T-learner*, "T" siendo una abreviación de *Two-Steps* en inglés. En la figura 1 se describe un diagrama con el proceso de entrenamiento y predicción de este algoritmo.

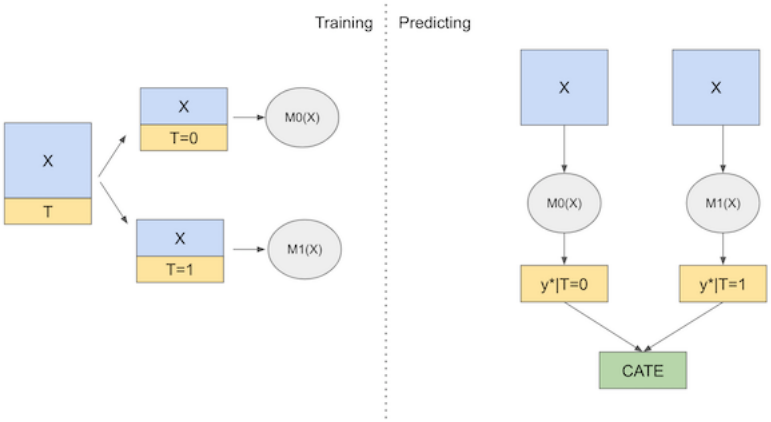

Fig. 1: T-Learner Diagram

Estrechamente relacionado con el *T-learner* está la idea de estimar el resultado usando todas las covariables (o *features*) y el indicador de tratamiento Z,

pero sin darle a éste último un rol fundamental en las predicciones. En este caso, el CATE predicho para un trabajador individual es entonces la diferencia entre los valores predichos cuando el indicador de asignación de tratamiento cambia, con todas las demás variables fijas. Nos referimos a este meta-algoritmo como *S-learner*, ya que utiliza un único estimador. En la figura 2 podemos ver su diagrama.

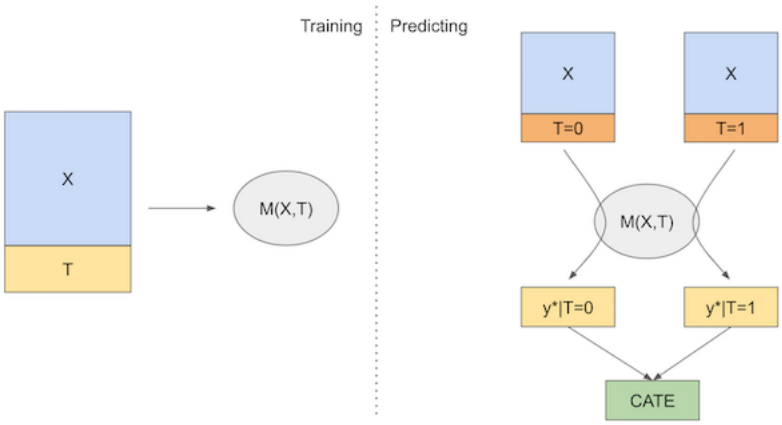

Fig. 2: S-Learner Diagram

Por último, hay un modelo de dos etapas llamado *X-learner* que en su primera etapa es idéntico al *T-learner*, luego en la segunda etapa imputamos el efecto del tratamiento para control y tratamiento usando los modelos de la primera etapa, finalmente calibra los modelos de la segunda etapa a través de un modelo de *propensity score*. En otras palabras, se favorece el modelo que fue entrenado utilizando más datos. En la figura 3 está el diagrama resumen.

Cabe destacar por cierto, que no todos los métodos que apuntan a capturar la heterogeneidad de los efectos del tratamiento se encasillan dentro de la clase de meta-algoritmos. Por ejemplo, algunos investigadores analizan la heterogeneidad por estimar los efectos promedio del tratamiento para subgrupos significativos (ver Hansen y Bowers, 2009 [8]). Otro ejemplo es utilizando *causal forests* (ver Wager y Athey, 2017 [26]). A pesar de que teóricamente el desempeño de los *X-learners* debiese ser superior a cualquier metodología supervisada, dada la naturaleza del problema, la calibración de hiper-parámetros y la calidad de los datos disponibles, podrían llevar a que empíricamente un modelo se comporte mejor o peor que otros en determinadas circunstancias. En la siguiente sección describiremos el proceso típico de regularización en *Machine Learning*, donde a través de *Grid Search* ajustamos los mejores hiper-parámetros para los meta-algoritmos, de manera de obtener las mejores estimaciones posibles en términos predictivos y evitar el problema de alta-varianza (*overfitting*).

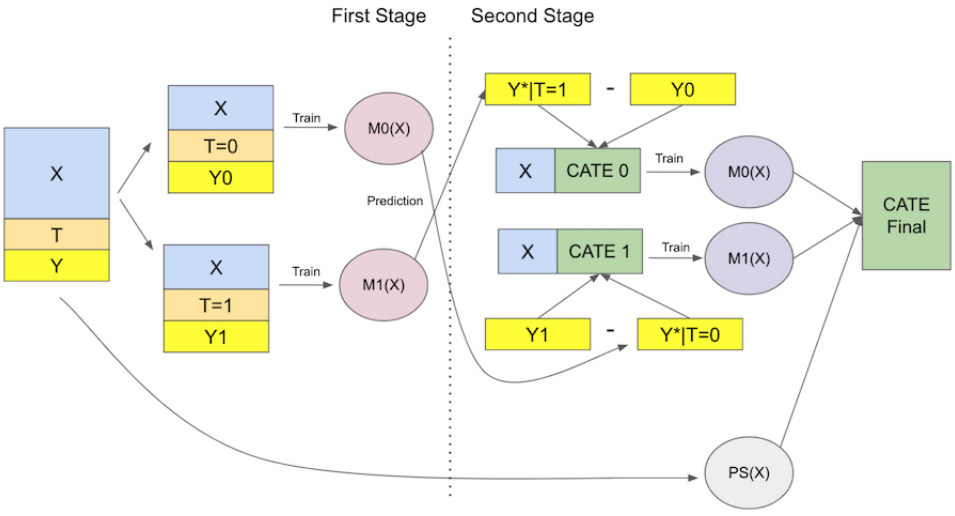

Fig. 3: X-Learner Diagram

Revisión de la Metodología

La flexibilidad de los estimadores de *Machine Learning* crea al mismo tiempo oportunidades y desafíos. Por ejemplo, es típico de que el número de parámetros del modelo que mejor se ajuste a un determinado problema de predicción, exceda el número de grados de libertad disponibles en la muestra. De hecho, en la literatura un apropiado método de regularización es un estándar requerido para mitigar el problema de *overfitting*. Más aún, la gran mayoría de modelos alcanzan su mejor ajuste solamente después de una configuración minuciosa de los hiper-parámetros que controlan el trade-off entre sesgo y varianza [4]. En particular, para la elección de los parámetros empleamos un algoritmo de fuerza bruta *Grid Search.* Por último, con el fin de validar un modelo correctamente ajustado, es una práctica estándar en *Machine Learning* realizar un proceso de división de la base de datos. A continuación, aplicaremos este *pipeline* para la estimación de los CATEs, procediendo a través de los siguientes pasos.

1. Dividiremos la base de datos $\mathcal{D}$ en dos particiones, un *set de entrenamiento* $\mathcal{D}_t$ y un *set de validación* $\mathcal{D}_v$, respectivamente para el ajuste de parámetros

[4] La regularización de los modelos de *Machine Learning* por lo general es un proceso de prueba y error empírica, en donde se suele automatizar el proceso de optimización de parámetros a través de algoritmos. Recientes trabajos han sometido a simulaciones exhaustivas distintos tipos de algoritmos de búsqueda tales como *Grid Search*, *Random Search* y *Genetic Algorithm* (Liashchynskyi, 2020) [18] donde en la mayoría de aplicaciones suele destacar *Grid Search* para efectos predictivos, a pesar de que en términos computacionales puede ser menos eficiente en comparación a lo demás algoritmos mencionados

y selección del modelo.

2. Ajustaremos los estimadores f_0, f_1 de los *outcomes* potenciales $Y(0)$ e $Y(1)$ en $\mathcal{D}_t$ y a través de un algoritmo de *Grid Search* escogeremos los hiperparámetros basados en el error obtenido en $\mathcal{D}_v$.

3. Computaremos el CATE, $\hat{\tau}_i := f_1(x_i) - f_0(x_i)$, para cada trabajador en $\mathcal{D}$ y ajustaremos un modelo interpretable a través del enfoque unificado SHAP (Lee, et al., 2017) [20] que describiremos más adelante.

Este *pipeline* nos permite encontrar los estimadores (tipo *caja negra*) con el mejor ajuste posible en el paso 2 sin tener en cuenta la interpretabilidad de sus predicciones. Al ajustar un modelo más simple e interpretable a los efectos imputados en el paso 3, podemos explicar las predicciones del modelo más complejo en términos de cantidades conocidas. Este procedimiento es particularmente adecuado cuando el tratamiento es una función más simple que la respuesta y también nos permite controlar lo granular del enfoque con el que estudiaremos la heterogeneidad.

Los datos de la EPS tienen una naturaleza de varios niveles; los trabajadores se agrupan en empresas e industrias y cada nivel está asociado con su propio conjunto de covariables. En este sentido, la literatura es rica con estudios de efectos causales en entornos multinivel, ver, por ejemplo, Gelman y Hill (2006) [2]. Sin embargo, este estudio está dirigido principalmente al estudios de los efectos de alto nivel (por ejemplo, nivel-industria), intervenciones en materia de bajo nivel (por ejemplo, a nivel-trabajador) y el aumento de la incertidumbre que viene con tal análisis. En las siguientes secciones describiremos cada paso de nuestra metodología en detalle.

Paso 1. División de $\mathcal{D}$: Para permitir la estimación imparcial del error de predicción y seleccionar parámetros de ajuste, dividimos el conjunto de datos $\mathcal{D}$ en dos partes con el 70% de los datos utilizados para el conjunto de entrenamiento $\mathcal{D}_t$ y el 30% para el conjunto de validación $\mathcal{D}_v$. Para prevenir *overfitting* en alguna variable, siguiendo a Johansson (2018) [9] equilibramos $\mathcal{D}_t$ y $\mathcal{D}_v$ construyendo un gran número de divisiones distribuidas de manera uniforme y seleccionando la que minimiza la distancia euclidiana entre el primer y segundo momento de todas las covariables. Aumentamos la influencia de la variable de tratamiento Z por un factor de 10 en esta comparación para asegurar que los grupos de tratamiento se dividen uniformemente entre $\mathcal{D}_t$ y $\mathcal{D}_v$.

Paso 2. Estimación de *outcomes* potenciales: El efecto de tratamiento promedio condicional es la diferencia entre los outcomes potenciales esperados. Aquí denotados μ_0 y μ_1. Bajo ignorabilidad escrito con respecto a X (ver arriba), tenemos que:

$$\mu_z(x) := E[Y(t) \mid X = x] = E[Y \mid X = x, T = t] \; para \; t \in \{0, 1\}$$

y por lo tanto, $\tau(x) = \mu_1(x) - \mu_0(x)$. Una ruta directa para estimar τ es independientemente ajustar los condicionales $E[Y \mid X = x, T = t]$ para cada valor de $t \in \{0, 1\}$ y computar su diferencia. Esto ha denominado recientemente el enfoque de meta-algoritmos para distinguirlo de otros paradigmas de aprendizaje (Künzel et al., 2017) [23]. A continuación, cubriremos brevemente la teoría que motiva este método y señalar algunas de sus deficiencias. Para estudiar la heterogeneidad, consideramos varios estimadores "meta" tomando en cuenta distintas bases, tales como algoritmos *Gradient Boosting*, regresiones *Ridge Lasso*, *Random Forest* y *Generalized Forest* [5].

Aproximamos μ_0, μ_1 usando las hipótesis f_0, f_1 y medimos su calidad por alguna métrica de precisión. Los riesgos condicionales de grupo empíricos y esperados se definen de la siguiente manera:

$$\underbrace{\mathcal{R}_t(f_t) := E[(\mu_t(x) - f_z(x))^2 \mid T = t]}_{\text{Riego esperado condicional de grupo}} \tag{2}$$

$$\underbrace{\hat{\mathcal{R}}(f_t) := \frac{1}{|G_t|} \sum_{i \in G_t} (f(x_i; \theta) - y_i)^2}_{\text{Riesgo empírico condicional de grupo}} \tag{3}$$

Nunca observamos μ_t directamente, sino que los algoritmos aprenden de observaciones ruidosas de y. La teoría del aprendizaje estadístico ayuda a resolver este problema al acotar el riesgo esperado en términos de su contra parte empírica y una medida de la complejidad de la función (Vapnik, 1999) [25]. Para distintas hipótesis en una clase $\mathcal{F}$ con una medida de complejidad particular $\mathcal{C}_{\mathcal{F}}(\delta, n)$ con dependencia logarítmica en n, se cumple con una probabilidad mayor que $1 - \delta$, lo siguiente

$$\forall f_t \in \mathcal{F} : \mathcal{R}_t(f_t) \leq \hat{\mathcal{R}}_t(f_t) + \frac{\mathcal{C}_{\mathcal{F}}(\delta, n)}{\sqrt{n}} - \sigma_Y^2,$$

donde σ_Y^2 es un límite de la varianza esperada en Y (ver Johansson et al, 2018 [9] para una derivación completa). Esta clase de límites ilustra el equilibrio entre sesgo-varianza que es típico en *Machile Learning* y motiva el uso de regularización para controlar la complejidad del modelo. Como se mencionó, en nuestra investigación, consideramos varios modelos *meta-learners* que estiman cada resultado potencial de forma independiente utilizando la minimizaciòn empírica del riesgo regularizada, resolviendo el siguiente problema.

[5] Este tipo de algoritmos difieren de los tradicionales basados en árboles de decisión, básicamente en que emplean un kernel ponderado, más no la versión clásica para así mitigar la *maldición de la dimensionalidad*, ver (Athey et al., 2019) [21]

$$f_t = \arg \min_{f(\cdot;\theta)\in\mathcal{F}} \hat{\mathcal{R}}_t(f(x;\theta)) + \lambda r(\theta) \tag{4}$$

Aquí $f(x;\theta)$ es una función parametrizada por θ y $r(\theta)$ es un modelo regularizador de parámetros como lo son los penalizadores l_1 (*Lasso*) y l_2 (*Ridge*).

A estas alturas cabe destacar una limitación que tienen los meta-algoritmos. En particular, se tiene que estos algoritmos no comparten información entre los diferentes *outcomes* potenciales. En problemas donde $Y(0)$ es una función más compleja que el efecto τ en sí mismo, los *meta-learners* implican un desperdicio en términos de poder estadístico (Künzel et al., 2017; Nie y Wager, 2017) [23] [16]. Como alternativa se podría adoptar un enfoque de redes neuronales TARNet de Shalit et al., 2017 [24] donde la arquitectura planteada permite aprender una representación de ambos grupos de tratamiento de manera conjunta, pero predecir potenciales *outcomes* separadamente. Esto tiene la ventaja de compartir información entre grupos de tratamiento en el aprendizaje del *outcome* promedio, pero permite flexibilidad en la aproximación del efecto.

Ahora considerando que, las métricas de riesgo y el problema de minimización descrito, están definidos con respecto a asignaciones observables del tratamiento, para estimar el CATE, queremos que nuestras estimaciones de los *outcomes* potenciales, de igual forma, sean precisas para las asignaciones del contrafactual. En otras palabras, queremos que el riesgo en toda la muestra,

$$\mathcal{R}(f_t) := E[(\mu_t(x) - f_z(x))^2]$$

sea pequeño. Cuando los grupos de tratamiento $p(X \mid T = 0)$ y $p(X \mid T = 1)$ están desbalanceados, el riesgo esperado dentro de un grupo de tratamiento puede no ser representativo del riesgo en la población completa. Este es otro inconveniente de los estimadores *meta-learners*, dado que no se ajustan a esta posible discrepancia.

En un trabajo reciente, Shalit et al., 2017 [24] caracterizan la diferencia entre $\mathcal{R}(f_t)$ y $\mathcal{R}_t(f_t)$ y acotaron el error en las estimaciones de CATE utilizando una métrica de distancia entre los grupos de tratamiento. En particular consideraron la familia de distancias IPM (*integral probability metric*), resultando en la siguiente relación entre el riesgo del grupo completo y del grupo de tratamiento.

$$\mathcal{R}(f_t) \leq \mathcal{R}_t(f_t) + IPM_{\mathcal{G}}(p(X \mid T = 0), p(X \mid T = 1)) \tag{5}$$

En Shalit et al., 2017 [24] los autores utilizaron la distancia media máxima basada en kernel (Ver Greton et al., 2012) [4] para regularizar y equilibrar las representaciones aprendidas por la arquitectura TARNet, minimizando el límite superior en el riesgo del CATE. En Johansson, 2018 [9] el autor utilizó esta misma arquitectura comparándola con meta-algoritmos, donde en general el desempeño fue bastante similar entre los modelos debido a un buen balance de las covariables. En particular, queda como una arista de investigación a futuro aplicar estas

metodologías en nuestro problema, sin embargo, como veremos en secciones más adelante, el balance de nuestras covariables es ideal, teniendo que las diferencias entre las distribuciones condicionales del tratamiento, son pequeñas.

Paso 3. Caracterización de la heterogeneidad en los CATEs: Después de ajustar los modelos f_0,f_1 para cada outcome potencial, el CATE se computa para cada trabajador como $\hat{\tau}_i = f_1(x_i) - f_0(x_i)$. A diferencia de los regresores lineales, las predicciones de la mayoría de los estimadores de Machine Learning son difíciles de interpretar directamente a través de los parámetros del modelo. Por esta razón, a menudo estos modelos son considerados métodos *black box* (Lipton, 2016) [13]. Sin embargo, en el estudio de la heterogeneidad, es crucial caracterizar para qué trabajadores el efecto de una intervención es bajo y para quienes es alto. Para lograr esto, adoptamos la práctica común de la interpretación *post-hoc*: ajustar un modelo más simple e interpretable $h \in \mathcal{H}$ a los efectos imputados $\{\hat{\tau}_i\}$.

Para esto seguiremos el enfoque unificado SHAP (ver Lee, et al., 2017) [20] donde se describe que $h(x_i)$ puede ser una función de un solo atributo, como por ejemplo el tamaño de la empresa donde trabaja el individuo, promediando efectivamente sobre otras covariables. Esta suele ser una buena manera de descubrir tendencias globales en los datos, pero descuida las interacciones significativas entre las variables, así como el modelo lineal. Como alternativa más flexible, realizamos una etapa intermedia de detección de la heterogeneidad ajustando modelos *Causal Forest* para así encontrar las variables que mayor heterogeneidad inducen al modelo, con todo lo demás constante (ver Johansson, 2018 [9]).

3 Resultados

A continuación, presentamos los resultados de nuestro análisis.

3.1 Balance de Covariables

Primero, investigamos si se cumple el supuesto de positividad comparando las estadísticas de las covariables de los grupos de tratamiento y control. En la figura 4, visualizamos las distribuciones marginales de cada covariable. Por otro lado, en la figura 5 observamos una proyección 2D t-SNE de todo el set de covariables (Maaten y Hinton, 2008) [14]. La diferencia observada entre las distribuciones marginales de los dos grupos de tratamiento es en general pequeña. También la proyección no lineal t-SNE revela poca diferencia entre los grupos de tratamiento. Cuanto menor sea el desequilibrio entre grupos de tratamiento, más cerca está nuestro problema de poder resolverse con aprendizaje supervisado estándar. Dicho de otra manera, el ratio de densidad $p(T = 1 \mid X)/p(T = 0 \mid X)$ es cercano a 1 y la distancia IPM entre distribuciones condicionales es pequeña (ver 5). Por lo tanto, no se espera que una arquitectura como la propuesta por Shalit et al., 2017 [24] tenga un gran efecto en los resultados (Johansson, 2018) [9].

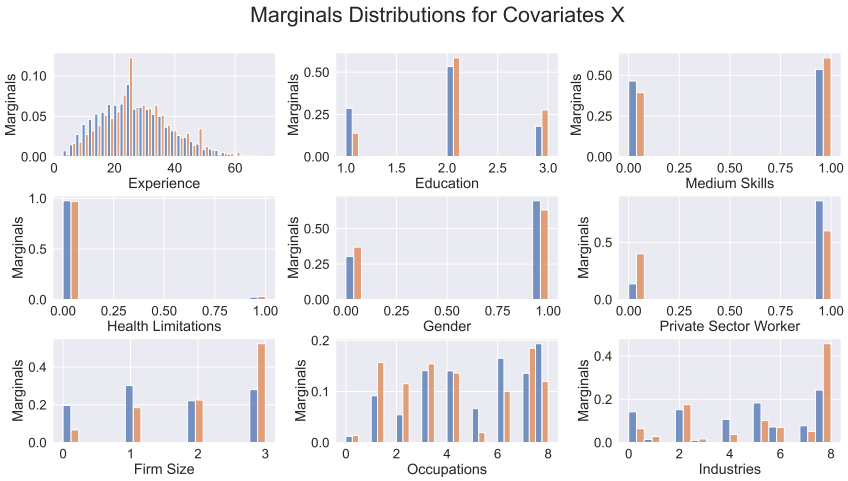

Fig. 4: Marginals of Covariates X. El color azul corresponden a los trabajadores del grupo de control, mientras que el naranjo a los de tratamiento (sindicalizados).

Fig. 5: 2D t-SNE projection of Covariates X

3.2 Estimación de los outcomes potenciales

En nuestro análisis, comparamos los tres meta-algoritmos presentados en el apartado metodológico. En principio, la primera etapa de cada uno de los modelos *off-the-shelf* estuvo basada en algoritmos de aprendizaje de *Gradient Boosting*, Regresiones *Ridge* *Lasso*, *Random Forest* y *Generalized Forest*. Para cada familia de estimadores, ajustamos modelos de ambos *outcomes* potenciales en el conjunto de entrenamiento $\mathcal{D}_t$ y la selección de parámetros (y modelos de primera etapa) a través *Grid Search* empleando el R^2 como métrica de validación en el set $\mathcal{D}_v$. Del proceso de búsqueda, el algoritmo de primera etapa ganador fue un Random Forest, modelo del que computamos nuestra estimación final (con sus hiper-parámetros respectivos ya ajustados). Para estimar incertidumbre en las

predicciones del modelo, realizamos *Bag of Little Bootstrapping* a nivel trabajador del conjunto de entrenamiento (Kleiner et al., 2012) [3] ajustando los modelos a cada muestra del *bootstrap*. Además, comparamos los meta-algoritmos con modelos de Propensity Score Matching y MCO Ajustados, metodologías clásicas de econometría.

Estimator	**ATE**	R^2
Naive	0.24	-
T-Learner	0.0707 [0.0514 - 0.09]	0.47
S-Learner	0.036 [0.0302 - 0.0418]	0.48
X-Learner	0.0304 [0.0175 - 0.0432]	0.47
PS Match	0.1085 [0.0461 - 0.1708]	0.09
OLS	0.0984 [0.0441 - 0.1526]	0.05

Table 1: Outcome Estimation

En la tabla 1, mostramos la estimación del efecto de tratamiento promedio (ATE) de cada modelo, el R^2 para los *outcomes*observables y los intervalos de confianza del 95% basados en el *bootstrap* empírico para los algoritmos *meta-learners*. Además, reportamos la estimación *naive* del ATE, es decir, la diferencia entre los *outcomes* promedios observados en los dos grupos de tratamiento. De los resultados, observamos que los *meta-learners* cuentan con un poder predictivo similar, mientras que las metodologías de econometría se quedan bastante por detrás en términos del R^2. En particular, el algoritmo *S-Learner* es aquel con mayor poder predictivo y reporta un valor del ATE igual a 3.6%, lo que es un resultado robusto de acuerdo a magnitudes encontradas anteriormente en la literatura. Landerretche, et al., 2013 [17], empleando un modelo de dinámico de dos etapas sobre la misma EPS, encontraron un ATE de 4.3% en salarios producto de la sindicalización. Más aún, el *X-Learner* alcanza este valor de ATE dentro de su intervalo de confianza empírico. Finalmente, en la figura 6 vemos las estimaciones de los CATE marginales a cada una de las covariables del problema. En general se aprecia bastante heterogeneidad en las primas salariales, pudiendo tener valores negativos en algunas industrias y positivos en otras. Similar ocurre con la variable categórica de ocupación y tamaño de la empresa.

3.3 Heterogeneidad en efectos causales

En esta sección examinamos más a fondo el CATE para cada trabajador imputado por el modelo de mejor ajuste, en este caso el *S-Learner* basado en *Random Forest*. En la sección anterior vimos una marcada heterogeneidad en la distintas covariables del problema, teniéndose en promedio un efecto del 3.6% sobre los salarios producto de la sindicalización.

Luego, para descubrir los canalizadores de la heterogeneidad, siguiendo a Johansson, 2018 [9] ajustamos un estimador de Causal Forest a los efectos imputados para así estimar la importancia de cada covariable, es decir, la frecuen-

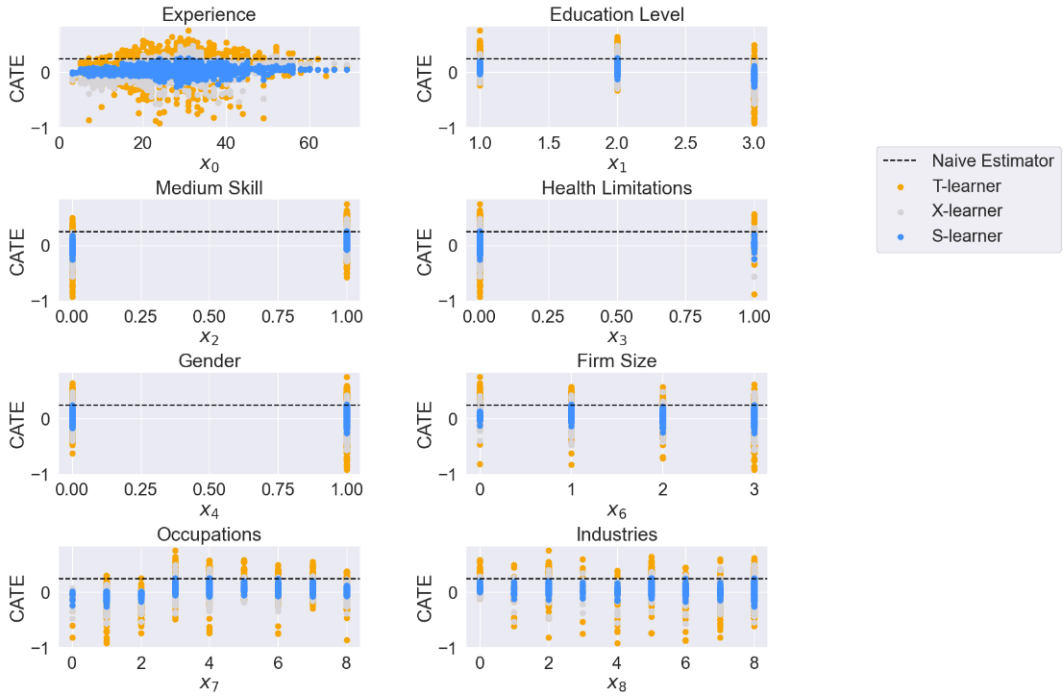

Fig. 6: Heterogeneidad en efectos causales para los distintos meta-learners.

cia con la que se utiliza para dividir los nodos de un árbol [6]. De acuerdo a esta metodología las covariables más importantes del Causal Forest fueron X_8 (industria), X_0 (experiencia), X_7 (ocupaciones), X_6 (tamaño de la firma) y X_4 (sexo del trabajador). En la figura 7 estratificamos el CATE imputado con respecto a estas covariables.

Según lo que se observa, cuando el trabajador tiene relativamente poca experiencia laboral, el premio salarial estaría contenido entre -30% y 3%, lo que sería menor que el ATE manteniéndose así hasta cumplidos 15 años. A los 40 años de experiencia, se observa que el premio por riesgo es el máximo en torno a 60%, para luego converger al valor medio. En el caso de profesionales, técnicos y administrativos, el premio sería menor que el ATE, teniéndose magnitudes entre -5 y -15% , mientras que para el resto de ocupaciones sería mayor, con niveles en torno al 20%. Por industria, se observa que solamente en manufacturas el efecto es menor al ATE con primas salariales de -5%, en minería se tiene el valor máximo entre las industrias con primas salariales cercanas al 40% y por su parte, servicios tiene valores bastante cercanos a la media. A nivel de género, se aprecia una brecha salarial de aproximadamente 8% en el tratamiento. De esta manera confirmamos la hipótesis de este estudio que buscaba identificar patrones no lineales en la heterogeneidad de los efectos causales en salarios dada

[6] para más detalles del árbol, sus nodos y del ranking de importancia por heterogeneidad, véase en el anexo la figura

la sindicalización de los trabajadores.

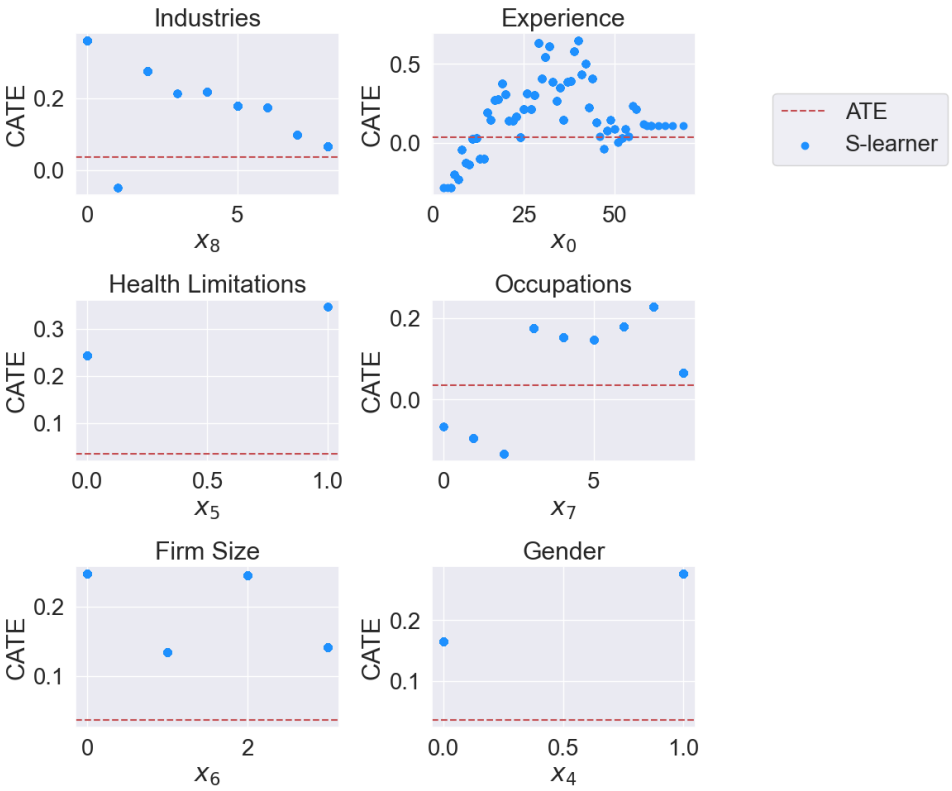

Fig. 7: Heterogeneidad en efectos causales estimada usando S-Learner.

3.4 Interpretabilidad del Modelo

En este trabajo sacrificamos la facilidad explicativa de los modelos lineales en virtud de la precisión en nuestras estimaciones y en poder mostrar patrones no lineales en los efectos del tratamiento. Aún así, nos gustaría comprender los determinantes del modelo y poder entender qué lleva al modelo a realizar una determinada predicción, más allá de encontrar la variables que inducen heterogeneidad.

La creciente tensión entre la precisión y la interpretabilidad de las predicciones de los modelos de *Machine Learning* ha motivado el desarrollo de métodos que ayuden a los usuarios a interpretar las predicciones. El marco SHAP propuesto por Lee y Lundberg, 2017 [20] identifica la clase de métodos de importancia aditiva de las covariables y muestra que hay una única solución de teoría de jue-

gos en esta clase que se adhiere a las propiedades deseables de consistencia, ausencia y precisión local. Esta metodología busca computar un modelo aditivo sencillo e interpretable h que hace *mapping* del modelo predictivo real. El hilo de unificación metodológica que teje SHAP a través de la literatura es una señal alentadora de que los principios comunes sobre la interpretación de modelos pueden llevar al desarrollo de nuevos métodos futuros.

Fig. 8: Shapley Additive Explanation Values.

En la figura 8 encontramos los valores SHAP para nuestro caso. De acuerdo a lo observado, podemos destacar que los valores más bajo de la *mapping-covariable* "mediana habilidad" (*skill*) disminuyen la prima salarian predicha dado el tratamiento, mientras que valores altos la aumentan. En el caso de la ocupación solamente tenemos claro que cuando la ocupación del individuo es profesional, técnico o administrativo, esta variable no aumenta el *outcome* dado el tratamiento. Por último, la última relación importante es respecto cuando la escolaridad es baja, en este caso las primas salariales tienden a disminuir.

4 Conclusion y discusión

El *Machine Learning* ofrece una amplia gama de herramientas para la aproximación de funciones y proporciona garantías teóricas para la estimación estadística en caso de, por ejemplo, errores de especificación del modelo. Esto, convierte al aprendizaje supervisado en un marco adecuado para la estimación de los efectos causales no lineales, desequilibrados o con datos de alta dimensionalidad. Sin embargo, la flexibilidad viene con un precio, muchos métodos vienen con parámetros de ajuste que son difíciles de establecer para estimación causal

y requieren de mucho ejercicio empírico de prueba y error. La interpretabilidad de los modelos sufre también. Si bien, se han hecho progresos de forma independiente en cada uno de estos problemas, como el *Grid Search* y los valores SHAP, aún no ha surgido un conjunto estandarizado de herramientas que sea amplio consenso en la literatura de *Machine Learning*.

En particular, en el análisis de los datos de la EPS, la metodología que propusimos parece adecuada para estudiar positividad, *outcomes* potenciales y heterogeneidad en los efectos causales. Encontramos que el modelo que mejor ajustaba las predicciones fue el *S-Learner* basado en *Random Forest*, el cual nos dio como resultado un ATE igual a 3.6%, mientras que la heterogeneidad del efecto en la prima salarial está bien contenida entre $[-0.2, 0, 6]$. Descubrimos varios patrones de heterogeneidad, destacando los efectos dispares entre industrias, donde las primas salariales por sindicalización serían relativamente más bajas en manufactura y servicios.La experiencia del trabajador también reporta heterogeneidad; se observa que cuando se tiene poca experiencia el premio salarial es menor al ATE, esto hasta aproximadamente los 15 años de experiencia. Sin embargo, el análisis también abre algunas preguntas metodológicas. La naturaleza multinivel de las covariable no se tiene en cuenta en los modelos meta-learners. Por otro lado, la regularización de modelos aplicados a datos de nivel múltiples ha sido considerablemente menos estudiada que para datos de un solo nivel (Johansson, 2018) [9]. Adicionalmente, como es señalado por varios autores (Künzel et al., 2017; Nie y Wager, 2017) [23] [16], los meta-algoritmos pueden desperdiciar poder estadístico al no compara los grupos de tratamiento. Esto correspondería a una de las razones por las que enfoques TARNet o de redes neuronales podrían tener una pequeña ventaja en términos predictivos y sería un buen ejercicio a futuro replicar este estudio con dichas metodologías.

Con todo lo señalado, los economistas deberíamos abrazar estas nuevas herramientas e incorporarlas a la econometría, sobre todo cuando se trata de problemas relevantes de predicción. Por ejemplo, la primera etapa de variables instrumentales es básicamente una predicción, es directo inferir que podríamos corregir aún más el sesgo al incorporar esta herramienta en la instrumentalización. En este trabajo vimos como es que podemos descubrir patrones complejos de heterogeneidad en el efecto de un tratamiento y es fácilmente reproducible a cualquier programa. Es momento que la disciplina comience a enseñar estas metodologías de manera formal en los cursos de econometría de pregrado, de manera de dotar a los estudiantes de herramientas que están siendo cada vez más requeridas en el mercado laboral (Athey y Luca, 2019) [22].

References

1. Adith Swaminathan, T.J.: Learning from logged bandit feedback", International Conference on Machine Learning, "Counterfactual risk minimization (2015)
2. Andrew Gelman, J.H.: Data analysis using regression and multilevel/hierarchical models (2006), cambridge university press.

3. Ariel Kleiner, Ameet Tawalkar, P.S.M.J.: The big data bootstrap (2012)
4. Arthur Gretton, Karsten Borgwardt, M.J.B.S.A.S.: A kernel two-sample test (2012), journal of Machine Learning Research.
5. Brian K Lee, Justin Lessler, E.A.S.: (2011), weight trimming and propensity score weighting
6. Contreras D., Hojman D., M.M.R.P.S.N.: "The impact of commuting time over educational achievement, A machine learning approach" (2018), serie de Documentos de Trabajo. Departamento de Economía - Universidad de Chile.
7. Donald B Rubin, N.T.: relating theory to practice, Biometrics, Matching using estimated propensity scores (1996)
8. J., H.B.: (2009), attributing effectsto a cluster randomized get out the vote campaign. Journal of the American Statistical Association, 104(487):873–885.
9. Johansson, F.D.: Machine learning analysis of heterogeneity in the effect of student mindset interventions (2018), atlantic Causal Inference Conference,1811.0597
10. Kausel E., Ventura S., V.M.D.D.V.F.: "does facial structure predict academic performance?" (2018)
11. Kononenko, I.: history, state of the art and perspective", Artificial Intelligence in medicine, "Machine learning for medical diagnosis (2001)
12. Leal F., M.C., E., Z.: "inflation forecast in Chile with Machine learning methods" (2020), documentos de trabajo. Banco Central de Chile.
13. Lipton, Z.C.: The mythos of model interpretability (2016)
14. Laurens van der Maaten, G.H.: Visualizing data using t-sne (2008), journal of machine learning
15. Mullainathan, S., Spiess, J.: "Machine Learning: An Applied Econometric Approach." Journal of Economic Perspectives p. 87–106 (2017)
16. Nie, X., Wager, S.: Learning objectives for treatment effect estimation (2017)
17. Oscar Landerretche, Nicolas Lillo, E.P.: A Two-Stage Approach Using Panel Data, The Union Effect on Wages in Chile (2013)
18. Petro Liashchynskyi, P.L.: Grid Search, Random Search, Genetic Algorithm (2020), journal of the American Statistical Association.
19. Rubin, D.B.: Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association **100**(469), 322–331 (2005)
20. Su-In Lee, S.L.: A unified approach to interpreting model predictions (2017), conference on Neural Information Processing Systems.
21. Susan Athey, J.T., Wager, S.: Generalized random forest (2019)
22. Susan Athey, M.L.: "economists (and economics(in tech companies", journal of economic perspectives (2019)
23. Sören R Künzel, Jasjeet S Sekhon, P.J.B., Yu, B.: Meta-learners for estimating heterogeneous treatment effects using machine learning (2019)
24. Uri Shalit, Fredrik D Johansson, D.S.: (2017), estimating individual treatment effect: generalization bounds and algorithms. In International Conference on Machine Learning
25. Vapnik, V.N.: An overview of statistical learning theory on neural networks (1999), iEEE transactions
26. Wager S, A.S.: Estimation and inference of heterogeneous treatment effects using random forests (2017), journal of the American Statistical Association.
27. Yann LeCun, Yoshua Bengio, G.H.: Deep learning. 436, nature, 521(7553) (2015)

5 Anexo

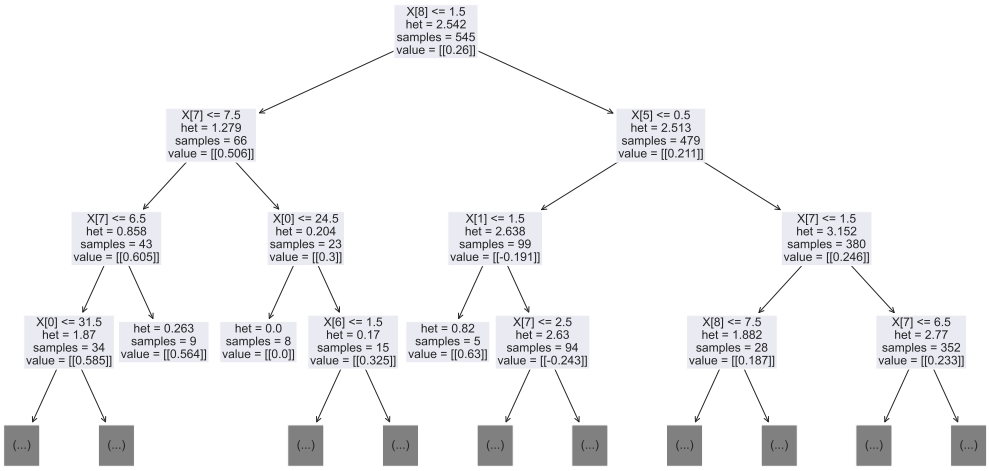

Fig. 9: Causal Forest Tree, first 3 nodes.

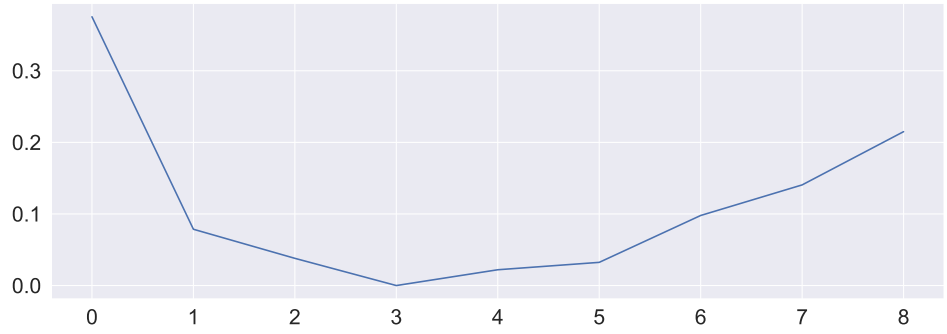

Fig. 10: Causal Forest Tree Feature Ranking.