

Efecto de la Educación Superior en México en el ingreso salarial de jóvenes profesionistas, utilizando Aprendizaje Automático Doble

Effect of Higher Education in Mexico on the Salary Income of Young Professionals, Using Double Machine Learning

Alejandro Pasos Ruíz¹, Emilio Gabriel Rejón Herrera²
Universidad Autónoma de Yucatán, Facultad de Matemáticas.
Periférico Norte, Tablaje Cat. 13615, Mérida, Yucatán, México.
¹apasos@correo.uady.mx, ²rherrera@correo.uady.mx.

Fecha de recepción: 30 de julio de 2025

Fecha de aceptación: 28 de agosto de 2025

Resumen. El objetivo de este trabajo, es utilizar inferencia causal para determinar si existe una relación entre el ingreso salarial y el nivel de educación superior. De tal forma, utilizando la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH), se extrajeron datos de jóvenes al incorporarse en el mercado laboral, con el fin de determinar dicha relación causa-efecto. Para resolver esto, se aplicaron herramientas de aprendizaje automático doble (Double Machine Learning, DML). Es así, que, con este esquema, se determinó que, contar con el nivel de educación superior, incrementa en un 26.1% dicho ingreso salarial. Por último, para proyectos futuros, se sugiere añadir otras variables y fuentes de información, así como ampliar el esquema propuesto.

Palabras Clave: Aprendizaje Automático Doble, Educación Superior, Ingreso Salarial, Jóvenes Profesionistas.

Summary. The objective of this work is to use causal inference to determine if there exists a relation between income and university education. Using the national survey for income and expenses (ENIGH), data about working young people were extracted with the goal of finding the mentioned relation. To achieve this, a double machine learning model was developed. In this way, this scheme determined that university education increases in 26.1% of the income. Finally, for future projects it is suggested to add more variables and other data sources and expand the proposed schema

Keywords: Machine Learning Double, University Education, Income, Young Workers.

1 Introducción

La educación superior se considera una de las principales alternativas, para mejorar la economía de los países y el ingreso salarial de su población. De tal forma, en México, se ha realizado un esfuerzo relevante para ampliar la matrícula y la oferta educativa, facilitando el acceso a la educación pública y privada. Por lo tanto, es importante determinar, si dicho esfuerzo por incrementar el número de egresados con título universitario, mejora sus condiciones económicas o, existen otros factores significativos (ya sea de tipo social y/o demográfico).

De tal manera, en este trabajo, se responde la pregunta: ¿cuál es el ingreso salarial de un joven profesionista, egresado de una institución de educación superior, en comparación con una persona que no accedió a una carrera universitaria, considerando la información obtenida, a partir de la encuesta ENIGH?. Asimismo, dicho cuestionamiento se puede aproximar, mediante el concepto de inferencia causal.

Es así, que, para resolver este tipo de problemas, se han utilizado diferentes modelos estadísticos, tales como las regresiones lineales o los modelos de evaluación A/B. En este caso, se contempló el uso de aprendizaje automático (Machine Learning, ML). Dentro de este campo, existe el aprendizaje supervisado, que permite establecer modelos predictivos para determinar el posible ingreso salarial de un profesionista, basándose en los factores antes mencionados. Sin embargo, estas predicciones pueden no ser suficientes para responder el planteamiento propuesto, ya que es necesario determinar la relación causa-efecto, entre las variables, educación superior y salario.

Por lo anterior, se propone aplicar aprendizaje automático doble (Double Machine Learning, DML), el cual permite estimar los efectos causales, de una manera robusta y escalable, entre las principales variables identificadas en la solución. Asimismo, se utilizaron los datos de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH), elaborado por el Instituto Nacional de Estadística, Geografía e Informática (INEGI). Dicha encuesta, tiene el propósito de identificar los diferentes aspectos socioeconómicos, así como obtener diversas métricas de la población.

2 Marco Teórico

Existen diversos estudios, que abordan el tema del ingreso económico en egresados de instituciones de nivel de educación superior; por ejemplo, en Tamborini (2015), se utilizaron datos sobre ingresos y se investigó la trayectoria laboral a través de 20 años, para estimar el efecto a largo plazo del incremento salarial. Por otro lado, Kim (2015), analizó el efecto de las distintas carreras, en la disparidad de ingresos. A su vez, la Organización para la Cooperación y Desarrollo Económico (OECD, 2020), publicó reportes sobre tasas de empleo y calidad de trabajos obtenidos por egresados. También, para estudios específicos de México (Levy, 2020), se concluye que, a pesar de la expansión del sector educativo, los ingresos se han visto afectados por una mala asignación entre empresas y sectores; lo que contribuye al estancamiento de los salarios y a la compresión de los rendimientos, para los trabajadores con mayor nivel educativo. Además, Cabrera-Hernández (2023), investigó las trayectorias académicas y proporciona un contexto de como estas, influyen en la incorporación al sector laboral y en los salarios. Estos estudios, se concentran en el ingreso de los egresados, pero no realizan un comparativo, con las personas que no cuentan con estudios de educación superior.

Por otro lado, tradicionalmente, los modelos económicos de la educación, identifican una ventaja salarial asociada al nivel educativo alcanzado. De tal manera, desde el modelo de Mincer (1974), se plantea una relación log-lineal entre ingresos y años de escolaridad. Es así, que, existen estudios empíricos recientes que utilizan diseños cuasi-experimentales, señalando, que la educación incrementa la productividad y, por lo tanto, los ingresos; de tal forma, el Instituto Mexicano de la Competitividad (IMCO), publicó un artículo donde analiza los ingresos promedio de egresados de diversas carreras universitarias, destacando que los mismos, en áreas como Ciencias, Matemáticas y Estadística, tienen ingresos significativamente superiores, en comparación con aquellos que solo han completado el bachillerato. Otro estudio de la OECD en 2025, presentó datos comparativos sobre los ingresos y tasas de empleo, entre jóvenes mexicanos con y sin, educación superior, mostrando una clara ventaja económica para quienes han completado estudios de nivel superior. También, Rodríguez (2025), estimó las tasas de retorno a la educación superior en México, evidenciando una diferencia significativa en los ingresos salariales entre egresados universitarios y aquellos, sin educación superior.

Sin embargo, estos estudios aplican principalmente, métodos estadísticos tradicionales lo que ocasiona ciertas limitaciones. En respuesta, se ha desarrollado métodos de inferencia causal, como el uso de variables instrumentales, diseños de regresión discontinua y, más recientemente, enfoques basados en aprendizaje automático. Asimismo, algunos estudios son: Card (2001), utilizó la proximidad geográfica a las universidades, como instrumento para estimar el ingreso; Heckman et al. (2006), se enfocó en la importancia de las habilidades no cognitivas y los factores de selección de las instituciones de educación superior; Athey y Imbens (2019), propusieron un modelo basado en árboles causales (Causal Forests), para modelar efectos del ingreso obtenido con o sin, contar con el nivel de educación superior.

Por lo anterior, este trabajo aplica los modelos de Aprendizaje Automático Doble, propuesto por Chernozhukov et al. (2018), para determinar inferencia causal. Es así, que el estudio original usa dichos modelos, para determinar el efecto causal que tienen los programas de entrenamiento, en los ingresos de los trabajadores. Este esquema, busca estimar el efecto de una variable de tratamiento (T) sobre una variable de salida (Y), utilizando variables de control o covariables (X). De tal manera, en los métodos tradicionales de inferencia causal, si las covariables tienen relaciones no lineales, el resultado se puede sesgar y no realizarse de manera correcta. El DML corrige este problema, al usar modelos de aprendizaje automático para identificar relaciones no lineales, entre las covariables.

Por lo tanto, el procedimiento consiste en dos fases:

1. Predicción de residuales: Se predice Y usando las covariables X, sin la variable de tratamiento. Esto permite obtener un residuo, que es la diferencia entre la predicción y el valor verdadero. De la misma forma, se predice T usando las covariables X, obteniendo su residuo.
2. Regresión de residuos: Se hace una regresión del residuo de Y, con el residuo de T. Con esto se obtiene el coeficiente efecto promedio de tratamiento (Average Treatment Effect, ATE), que representa el efecto causal de T sobre Y, libre de sesgo.

Este esquema tiene ventajas, tales como el uso de modelos de aprendizaje no paramétrico y los modelos de ensamble de árboles. También, al ser modelos no lineales, se reduce el sesgo por omisión de variables y se minimiza el ruido. Finalmente, se producen estimaciones consistentes, independientemente del número de covariables.

3 Diseño experimental

Para aplicar este modelo, se utilizaron datos del año 2022, de la encuesta nacional de ingresos y gastos de los hogares (ENIGH). De tal manera, se realizó un filtrado para incluir a las personas de entre 22 y 35 años de edad; además, para un mejor tratamiento, dichos datos fueron limpiados y escalados. Las variables de la encuesta seleccionadas manualmente son las siguientes:

- Tratamiento (T). Educación superior.
(variable de tipo binario, si tiene estudios | no tiene estudios).
- Resultado (Y). Ingreso laboral trimestral.
(variable de tipo numérico, se escaló para un mejor tratamiento usando una conversión logarítmica).
- Variables de control o covariables (X):
 - Edad.
 - Género.
 - Estado civil.
 - Tamaño del municipio.
 - Número de personas en el hogar.
 - Condiciones del hogar.
 - Ocupación y sector laboral.
 - Acceso a salud.
 - Acceso a internet o computadora.
 - Tenencia de vivienda.

4 Metodología

A partir de estas variables, se sigue el procedimiento de aprendizaje automático doble (DML). Para la primera fase, se realizó validación cruzada con 10 grupos y se calculó el promedio del resultado, con el fin de estimar Y, T. Se consideraron los siguientes algoritmos:

- a. Modelos basados en ensamble de árboles, tales como XGBoost y Random Forest.
- b. Regresión lineal con regularización Lasso.
- c. Máquinas de vector soporte.

Es así, que el objetivo de la inferencia causal es identificar las relaciones, más allá de las posibles correlaciones existentes en los datos. Cabe señalar, que, en este caso, dichas correlaciones, no son suficientes. Por lo tanto, al aplicar el aprendizaje automático doble, se obtiene el efecto promedio de tratamiento (Average Treatment Effect, ATE). Esto, al aplicar regresión lineal, a las dos variables residuales que se generaron a partir de la fase anterior. De tal forma, el coeficiente ATE indica el nivel de relación causal, que se tiene entre las variables Y, T.

5 Resultados

En la primera fase de aprendizaje automático doble, se generan modelos para predecir la variable de resultado (Y), basándose en las variables de control (X). Dado que, Y es una variable numérica, se tiene que tratar el problema como de regresión. Para esto, se utilizan las versiones de los algoritmos mencionados, como modelos de regresión. Al usar la validación cruzada, es necesario definir métricas para comparar los resultados de los algoritmos. Es así, que las métricas más utilizadas son R² y MSE. De tal manera, R² mide la proporción de la variabilidad del resultado, que el modelo puede explicar, con las variables predictoras, usando la siguiente formula:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Donde y_i es el valor real y $\hat{y}_i$ es la predicción. Adicionalmente, la métrica MSE mide qué tan lejos están (en promedio), las predicciones del modelo de los valores reales, penalizando los errores más grandes. La fórmula es:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Los resultados se muestran en la Tabla 1. Se puede observar que XGBoost, obtuvo las tasas más altas en R², así como las más bajas en MSE; por lo que constituye el mejor desempeño. Por lo tanto, este algoritmo se utilizó para elaborar el modelo final.

Tabla 1. Resultados para predecir la variable ingreso (Y), usando las variables de control (X).

Algoritmo	R ²	MSE
Random Forest	0.041296	2.529778
XGBoost	0.044377	2.521647
Lasso	-0.000071	2.638926
SVM	0.040234	2.562613

Para calcular el segundo residual, se busca predecir la variable de tratamiento T, usando las variables de control X. En este caso, T es una variable binaria que determina si una persona tiene o no, estudios de educación superior. De tal forma, se generó un modelo de clasificación, que busca predecir la probabilidad de que una persona tenga estudios de educación superior, utilizando las variables de control. Para este tipo de problemas, existen diferentes métricas para comparar el desempeño de los algoritmos. En este trabajo, se utilizan dos de las más comunes: 1) efectividad (promedio de predicciones correctas) y 2) área bajo la curva (AUC), que mide la probabilidad de que un clasificador, separe eficientemente los casos positivos de los negativos. En la tabla 2, se muestran los resultados, en donde se puede observar que el algoritmo de Random Forest obtuvo un mejor desempeño, por lo que dichas predicciones se utilizan, como parte del segundo residual.

Tabla 2. Resultados para predecir la variable de tratamiento (T), usando las variables de control (X).

Algoritmo	Efectividad	AUC
Random Forest	0.8308	0.8974
XGBoost	0.8096	0.8692
Lasso	0.7803	0.8235
SVM	0.8135	0.8441

Finalmente, los resultados de los modelos anteriores se utilizan como base para la segunda fase. Esta, realiza una regresión lineal comparando los residuales de ambos modelos. Además, el coeficiente ATE representa la relación causal entre Y, T. Por lo tanto, en la regresión lineal calculada se obtuvo un coeficiente de ATE = 0.233, con un error estándar de 0.0105. Dicho resultado, sugiere que, si existe una relación causal entre variables, al no ser cercano a cero. Sin embargo, no cuantifica el valor verdadero.

Para esto, hay que recordar que la variable Y se preprocesa, usando una escala logarítmica. Para tener un resultado en la escala original, se necesita de-escalar el resultado de ATE. Al realizar esto, obtenemos el valor original $e^{0.233-1}=0.261$. Lo cual significa que, contar con el nivel de educación superior, incrementa el ingreso salarial en un 26.1%. Al tener un error estándar bajo, se puede concluir que los resultados obtenidos, son estadísticamente válidos.

6 Conclusiones

Este trabajo busca resolver la pregunta: ¿Cuál es el ingreso de un joven profesionista, egresado de una institución de educación superior, en comparación con una persona que no accedió a una carrera universitaria?. Se realizó un estudio de trabajos similares y se propuso utilizar aprendizaje automático doble, el cual hasta el momento no se había utilizado para este problema. Para aplicar este modelo, se utilizaron los datos de la encuesta ENIGH del año 2022, entre los jóvenes de 22 a 35 años, con o sin, estudios universitarios en México. Para esto, se usaron algoritmos de XGBoost y Random Forest (para la primera fase) y, Regresión lineal (para la segunda);

de tal manera, se puede concluir, basados en la información proporcionada por el ENIGH, que el nivel de educación superior incrementa en un 26.1% los ingresos percibidos.

Dichos resultados, coinciden con estudios similares. Además, el uso de modelos de aprendizaje automático doble permite obtener resultados con una mayor exactitud y pueden ser escalables. Por otro lado, para proyectos futuros, se sugiere realizar un análisis más profundo de las variables de control, ya que en este trabajo se seleccionaron manualmente, por lo que un estudio más detallado o un análisis estadístico de las variables, podría ser importante para determinar las mejores posibles, así como identificar otras de tratamiento o, de resultados.

De igual forma, se podría aplicar el modelo a nivel regional con el fin de determinar si la relación causa efecto es distinta, de acuerdo con la posición geográfica. Lo anterior, puede ayudar a determinar si en algunas regiones no es tan relevante tener estudios de educación superior. Con esta información se podría llevar a realizar una investigación más profunda, con el fin de tomar mejores decisiones de cómo mejorar las instituciones de educación superior para que sus egresados accedan a mejores oportunidades. Además, permitiría identificar la importancia de las variables a nivel regional y observar si existe una variación entre estas. Esto es, debido a que cada región puede tener variables de control específicas. Asimismo, el ENIGH cuenta con datos bianuales de 1990 a 2024; para realizar un comparativo en el tiempo de la evolución de las relaciones causales y, determinar si las estrategias educativas están funcionando o no. Finalmente, identificar cuáles son los sectores en donde tener educación superior tiene una mayor incidencia en el ingreso salarial, podría coadyuvar en la implementación de mejores estrategias, para promover dichos sectores y aumentar la calidad de vida de más egresados.

Referencias

- [1] Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685-725.
- [2] Cabrera-Hernández, F., et al. (2023). Full-time schools and educational trajectories: Evidence from Mexico. *Economics of Education Review*
- [3] Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5), 1127-1160.
- [4] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.
- [5] Heckman, J. J., Stixrud, J., & Urzúa, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24(3), 411-482.
- [6] Instituto Mexicano para la Competitividad (IMCO). (2024). *Compara Carreras 2024: ¿Formamos el talento que México necesita?* Instituto Mexicano para la Competitividad
- [7] Kim, C., Tamborini, C. R., & Sakamoto, A. (2015). Field of study in college and lifetime earnings in the United States. *Proceedings of the National Academy of Sciences*, 112(6), 1500-1506.
- [8] Levy, S., & López-Calva, L. F. (2020). Persistent misallocation and the returns to education in Mexico. *The World Bank Economic Review*, 34(2), 284-311.
- [9] Mincer, J. (1974). *Schooling, Experience, and Earnings*. Columbia University Press.
- [10] OECD. (2020). *How are young graduates settling into the labour market?* Paris: OECD Publishing.
- [11] OECD. (2025). *Education at a Glance 2025: Mexico*. OECD Publishing
- [12] Rodríguez Galván, J. Y. (2025). Returns to higher education in Mexico: A cross-sectional analysis (2010-2024). *Economics*, 13(2), 43.
- [13] Tamborini, C. R., Kim, C., & Sakamoto, A. (2015). Education and lifetime earnings in the United States. *Demography*, 52(4), 1383-1407