GASTO EN EDUCACIÓN EN COLOMBIA

Daniel Franco

Samuel Malkún

Diego Osorio

Resumen

El gasto en educación es determinante para una economía al ser fundamental para el ingreso futuro. En el presente trabajo se desarrolla un modelo para predecir el gasto en educación a nivel hogar. Para ese objetivo se utilizan datos de la Encuesta Nacional de Calidad de Vida del DANE de 2021, que cuenta con información a nivel hogar e individuo. Se desarrollan distintos modelos de aprendizaje de máquinas y una métrica de evaluación que penaliza en una mayor medida la sobreestimación. El anterior proceso llevó a la escogencia de un modelo de Random Forest de regresión con 350 árboles y una profundidad de 20. El modelo muestra la importancia de los términos no lineales y la interacción entre variables para predecir el gasto en educación.

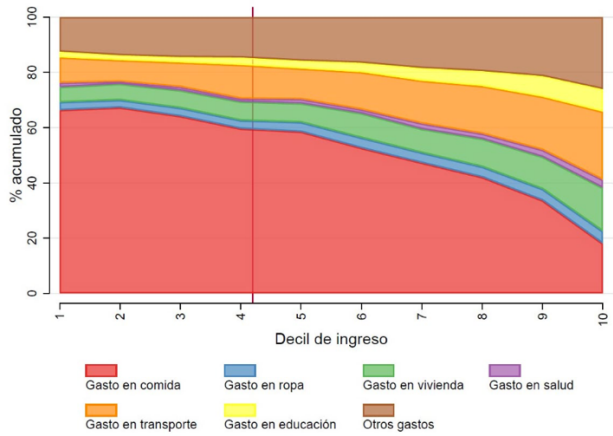
https://github.com/dfosorio111/BD-ML---FinalProject

Introducción

El gasto en educación es determinante para una economía, siendo el nivel de educación uno de los principales determinantes de ingreso. Si bien los gobiernos provén educación pública a los más pobres, los hogares aún tienen que enfrentar cierto costo asociado a la educación de sus hijos. En los hogares de menor ingreso el gasto en educación es aún más importante, pues la movilidad social va a depender en gran medida del logro educativo de los hijos.

En Colombia se estima que los hogares gastan cerca del 3,2% de sus ingresos en educación, según la última Encuesta Nacional de Calidad de Vida (ENCV) llevada a cabo por el DANE. Sin embargo, se observa que esa cifra no es homogénea entre grupos por ubicación y por nivel de ingreso. En efecto, los hogares en cabeceras gastan alrededor del 3,6% en educación, mientras que esa cifra desciende a 0,6% para los hogares en centros poblados rurales. La ENCV también muestra que el gasto en educación de los hogares no es el mismo para todas las regiones del país, siendo más bajo en la región Oriental y en San Andrés que en Bogotá. Por último, existen brechas por nivel de ingreso. Según el BID (2021), los hogares más pobres gastan una menor proporción de sus ingresos en la educación de sus hijos, lo que puede explicar brechas existentes en acceso y calidad que requieren estrategias para ser disminuidas.

Figura 1. Distribución del gasto de los hogares por decil de ingreso per-cápita mensual



Fuente: BID (2021)

El objetivo del presente proyecto es predecir la proporción que representa el gasto en educación con respecto al ingreso total a nivel hogar en Colombia por medio de distintas metodologías que permitan identificar los principales predictores y proveer recomendaciones de focalización de política pública. La intención es ubicar a aquellos hogares en donde el gasto en educación sea muy bajo para llevar registro de los logros educativos de los miembros de ese hogar y otorgar ayudas de ser necesario.

Por tanto, como se menciona en repetidas ocasiones en el presente documento, se considera que tiene mayor gravedad sobreestimar las predicciones, porque la intención principal es identificar aquellos hogares que están invirtiendo en educación por debajo de sus capacidades y buscar comprender las razones para formular políticas que los lleven a cambiar esta realidad.

Para predecir adecuadamente el gasto en educación se utilizará la Encuesta Nacional de Calidad de Vida del 2021, de la que se tienen microdatos frente a distintas variables que influyen en el nivel de gasto que un hogar destina a la educación. En cuanto a la metodología, se abordarán distintos modelos de regresión partiendo desde el más simple – regresión lineal – y llevándolo a métodos más complejos como bosques de decisión. En el presente documento se observa un destacado resultado de aquellos modelos que capturan no linealidades dejando en claro la importancia de las interacciones en el análisis. Como se verá más adelante, se logra obtener una predicción satisfactoria por medio de bosques aleatorios, al ser seleccionado como el mejor modelo bajo el criterio de decisión que penaliza con mayor severidad la sobreestimación. Posteriormente, un análisis de la relevancia de las variables muestra la importancia de complementar con un entorno propicio donde el acceso a tecnología, las características de la vivienda y el modo de transporte cuentan con un rol fundamental.

Revisión de Literatura

La importancia de analizar los determinantes del gasto de los hogares en educación surge de las conocidas ventajas en capital humano, movilidad social, ingresos y oportunidades futuras que provienen de un aumento en los años educativos y de una mejor calidad en la educación impartida (Banerjee y Duflo, 2011). Además, de acuerdo con Deaton (2003) una correcta medición de la pobreza no implica solo tener en cuenta mediciones sobre el ingreso de los hogares, sino también considerar los patrones de consumo, pues una medición complementaria de ingresos y gastos permite generar mejores indicadores en términos de pobreza, índices de precios de bienes y servicios y diferencias en costos de vida.

A priori, desde un análisis inferencial resulta claro que factores como el nivel de ingreso del hogar, el nivel educativo de los padres, el contexto socio económico, el tamaño de la familia, la ubicación geográfica, entre otros aspectos, son determinantes importantes del gasto en educación de los hogares (Banerjee y Duflo, 2011; Bayar y Ilhan, 2016; Chi y Quian, 2016).

Por su parte, Ahmad y Fatima (2011) muestran que el ingreso del hogar, los activos, el número de personas que genera ingresos en el hogar, el tamaño de la familia, la región y el sexo de la persona cabeza de hogar son variables significativas para clasificar el nivel de gasto de los hogares, haciendo uso de redes neuronales.

Para culminar, Gao et al (2020) muestran el alcance del *Machine Learning* para la focalización de políticas públicas. En Gao et al (2020) se utilizan *Random Forest* de clasificación para abordar la problemática de inseguridad alimentaria. Con lo anterior en mente, se considera relevante poder sacar provecho del alcance del aprendizaje de máquinas para la focalización de recursos para la educación. Por ejemplo, dada la evidencia de la efectividad de los *vouchers* educativos en Colombia (Angrist et al, 2002; Angrist et al, 2004), sería posible focalizar estas estrategias a aquellas familias que revelen mayor necesidad de apoyo para poder incrementar su gasto en educación.

Datos

Para el presente proyecto, se hará uso de la Encuesta Nacional de Calidad de Vida (ENCV) del 2021 que provee el DANE de forma pública. Se identificarán los módulos relevantes para

predecir el gasto de los hogares en educación y consecuentemente los principales predictores, a partir de la literatura y la observación empírica de los datos. Es importante mencionar que la encuesta cuenta con observaciones a nivel vivienda, a nivel hogar y a nivel individuo, por lo que las variables construidas y los resultados obtenidos serán a nivel hogar.

La variable dependiente consiste en la proporción del gasto en educación con respecto a los ingresos totales del hogar (Chi y Qian, 2016). En consecuencia, se trata de una variable continua en el rango de 0 a 1. Para construirla es necesario agregar los distintos gastos educativos a nivel hogar que se registran en la ENCV (tanto dentro de la escuela como fuera de ella) y compararlos con su ingreso.

Tabla 1: Estadísticas del gasto en educación (% del total)

	Promedio	Mediana	Varianza	Mín.	Máx.
Gasto en educación	0.03028	0.01120	0.01705	1.117886e-06	6.5
Logaritmo del gasto en educación	-4.525272	-4.492161	2.139035	-13.70407	1.871802

A priori, la literatura sugiere que el nivel educativo de los padres, el nivel socioeconómico, la ubicación geográfica, el sexo de los individuos, el número de personas del hogar, la edad de las personas que se están educando, entre otras variables, son predictores relevantes. Las anteriores variables se encuentran en la ENCV, en conjunto con posibles predictores adicionales a nivel hogar y otros a nivel individuo que permiten la construcción de variables a nivel hogar.

Por consiguiente, se requiere una exploración exhaustiva de la base de datos y la construcción de múltiples predictores que lleven desde la información original disponible, hacia una base consolidada con predictores acordes a lo encontrado en la literatura y al contexto nacional. A continuación, se muestran estadísticas y gráficos preliminares de algunas de las variables de la base frente al porcentaje de gasto en educación. En todos los casos se toma el logaritmo de ese porcentaje, debido a que la distribución se acumula en la cola izquierda con unos cuantos atípicos elevados.

Hay una relación positiva entre el gasto en educación de un hogar y la proporción de estudiantes en el mismo. En cuanto al tiempo de transporte a la institución educativa, observamos que la relación es menos evidente y depende del medio de transporte utilizado. Parece haber una relación positiva para todos los medios de transporte, exceptuando quienes se desplazan a la institución educativa en caballo o mula. El hecho de que la relación no sea clara para cualquier medio de transporte da pistas de la importancia de incluir interacciones en el modelo (ver Figuras 2a y 2b en el Anexo)

A priori, se podría pensar que hay una relación entre el máximo nivel educativo y el gasto en educación, pues los jefes de hogar con mayor nivel educativo reconocen la importancia del mismo y destinan una mayor parte de su ingreso a ese rubro de gasto. Sin embargo, no se observa una relación clara en los datos. Es posible que el mecanismo anteriormente descrito se compense en el hecho de que los hogares con mayor educación tienen mayor ingreso y gastan, relativamente, un porcentaje menor de su ingreso en educación. Por último, no se observan

diferencias notorias entre los hogares que se autodenominan como pobres o que tienen computador (ver Figuras 3a y 3b en el Anexo).

Metodología

El problema de predecir la proporción de gasto a nivel hogar es continuo, lo que implica una regresión. Partiendo de modelos de regresión lineal simple para determinar un ajuste preliminar de los datos, se implementarán modelos de mayor complejidad en el caso de regresión que involucren técnicas de regularización como lo son Lasso (penalización L1), Ridge (penalización L2) y Elastic Net para penalizar la complejidad de los modelos en función de los predictores.

Paralelamente se desarrollarán modelos de ensamble que exploran las interacciones no lineales como Arboles de Decisión para Regresión, Bosques de Decisión, Random Forest Regressor, técnicas de Resampeo (Bagging de Regresión), modelos con aprendizaje exhaustivo como Gradient Boosting y XGBoost, así como arquitecturas de mayor complejidad con mayor poder predictivo a expensas en interpretabilidad de sus predictores y ajuste de hiper parámetros como *super-learners*. Cabe resaltar que los modelos presentados en este informe fueron escritos en código de R y Python.

Tanto en el caso de los modelos lineales como no lineales, los hiperparámetros óptimos se sintonizan mediante técnicas de validación cruzada para determinar los parámetros de modelo que penalicen con mayor severidad la sobreestimación, ya que es más relevante identificar los hogares en los que el gasto en educación es muy bajo.

En caso tal de que se lleve a cabo alguna política con base en un umbral del gasto en educación, será mejor incluir a hogares que no necesitan una ayuda a excluir a quienes si lo necesitan.

Para escoger el mejor modelo se utilizarán dos métricas de evaluación: RMSE y la función personalizada (custom) que se define a continuación: en caso de que el valor predicho de la proporción del gasto en educación sea mayor al valor real, entonces se penaliza al cuadrado, ya que queremos que se penalice con mayor severidad el sobreajuste. De lo contrario se penaliza con el valor absoluto del error.

$$Función\ error = \begin{cases} error^2 & si\ error > 0 \\ |error| & d.l.c \end{cases}$$

Definiendo:

$$error = y_{predict} - y_{real}$$

En la Tabla 2 en los Anexos se presenta el carrete de 15 especificaciones realizados para la investigación del proyecto donde se aprecia el tipo de modelo, la descripción preliminar de sus hiperparámetros, RMSE y la función personalizada descrita anteriormente para cada modelo.

A partir de los resultados obtenidos en la Tabla 2 en la sección de Anexos, se observa que el mejor a partir de las métricas objetivo de la función personalizada y el RMSE es un modelo de ensamble Random Forest de Regresión con los siguientes hiperparámetros: Random Forest de

Regresión con 350 árboles, con profundidad máxima de árbol de 20, que utiliza 40 variables en cada nodo para determinar el Split, con un mínimo de observaciones en un nodo para considerar hacer Split de 60. Este modelo fue entrenado para minimizar la métrica ('scoring') RMSE, se obtuvo un valor de la función personalizada de 0.0126 y RMSE de 0.0340

Para realizar la sintonización de los hiperparámetros se creó un objeto de la clase GridSearchCV de Python (librería sklearn) que construye una grilla de parámetros sobre las cuales realiza la sintonización exhaustiva mediante todas las posibles combinaciones de los parámetros. Sobre los parámetros del modelo la profundidad del arbol y el mínimo número de observaciones para hacer Split permiten controlar la complejidad del modelo. Estos parámetros son reportados por la comunidad por heurística que tienen un comportamiento eficiente para rangos entre $[7 - 20]$ mientras que el segundo por lo general es entre el 0.5 y 1% de las observaciones totales de los datos de entrenamiento. Por otro lado, el parámetro de variables en cada nodo para determinar el Split se determina normalmente como $\sqrt{variables}$ pero se recomienda utilizar hasta el 30-40% de las variables totales en la base de entrenamiento, entonces se escogieron 5 valores enteros entre 10 y 40. Finalmente el número de árboles se definió entre 3 valores entre 100 y 600 árboles en la grilla.

Los hiperparámetros descritos anteriormente presentan una ventaja desde el punto de vista de Computer Science y algoritmos de ajuste del modelo porque permiten implementar medidas de control de complejidad de modelo, control sobre el sobreajuste, así como también permiten alcanzar una especificación sobre la aplicación en cuestión en relación con la métrica de evaluación que ajusta el modelo, a diferencia de otras plataformas.

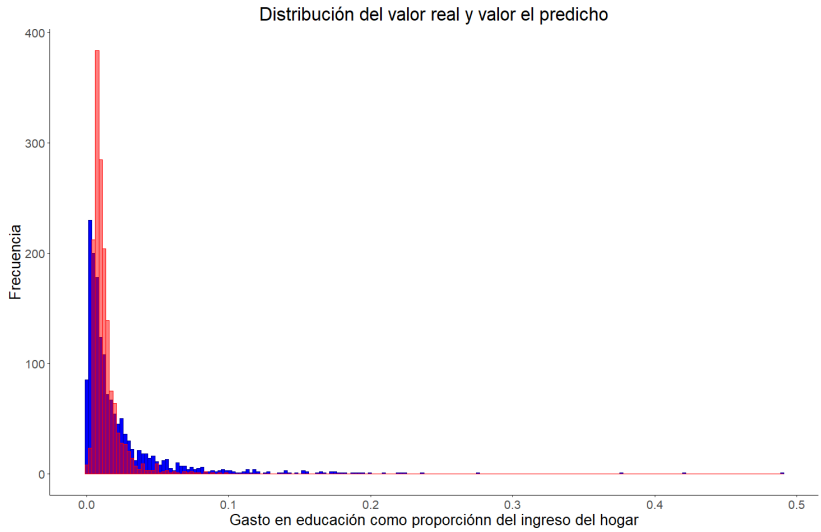
El modelo de Random Forest de Regresión es un modelo de ensamble de múltiples arboles de decisión con Bagging (resampleo con remplazo sobre la base de datos) que para solucionar el problema de estimar n arboles muy similares entre sí, en el Split de cada nodo solo utiliza p predictores para poder estimar diferentes partes de la distribución de los datos de manera más eficiente y generando modelos más robustos. Este modelo aplica la técnica de resampleo de Bootstraping sobre los árboles que ajusta, que estos árboles son ajustados con muestras que son tomadas de la base de datos de entrenamiento con reemplazo, los cuales son calculados para obtener el promedio de la métrica que se desea optimizar.

Resultados

Teniendo en cuenta lo mencionado en la sección anterior, en el presente aparatado se presentan los principales resultados de las predicciones realizadas.

En primer lugar, en la figura 4 se observa de color azul la distribución de los datos reales de la base de testeo y en color rojo los resultados predichos. Como se puede apreciar, la función de distribución es bastante similar en ambos casos, aunque los datos predichos se encuentran más agrupados que los datos reales, lo que resalta la dispersión existente en la proporción del ingreso de los hogares destinado en gastos en educación, con un buen número de valores que podrían considerarse como atípicos.

Figura 4: valores reales y predichos



Lo que se observa gráficamente en la figura anterior, se puede corroborar a continuación en la Tabla 3. Las predicciones son similares para los resultados más bajos, pero en el tercer cuartil y valores superiores las cifras se empiezan a alejar.

Tabla 3

Datos	Mínimo	25% (Q1)	50% (Q2)	75% (Q3)	Máximo	Promedio	Desviación
Reales	0.0002%	0.4665%	1.0396%	2.4390%	48.9583%	2.2458%	3.6528%
Predichos	0.0394%	0.7096%	0.9780%	1.4613%	9.7314%	1.2807%	1.4613%

Lo anterior permite intuir que la subestimación es más notoria que la sobreestimación, como se ratifica en la Figura 5 y en la Tabla 4. Acorde con la función objetivo planteada en el presente trabajo es menos deseable sobreestimar, por lo que los resultados obtenidos son satisfactorios.

Figura 5: Distribución del error

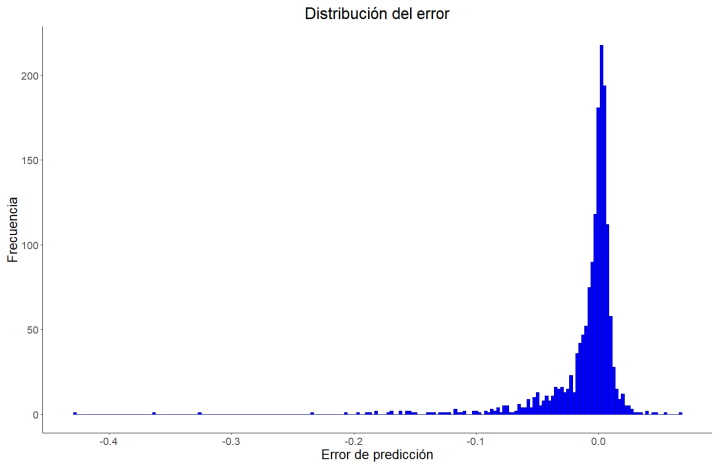


Tabla 4

Mínimo	25% (Q1)	50% (Q2)	75% (Q3)	Máximo	Promedio	Desviación
-42.69%	-1.12%	-0.02%	0.43%	6.84%	-0.97%	3.26%

De igual forma, resulta interesante contemplar las variables más importantes del modelo seleccionado. En la Figura 6 se confirma la relevancia de las variables que denotan el nivel de ingreso de los hogares, como se señaló en la revisión de literatura. Además, la proporción de trabajadores o estudiantes en el hogar es relevante. Las anteriores variables de ingreso y proporción de personas que estudian tienen una relación directa bastante clara con la variable que se busca predecir.

Figura 6: Importancia de las variables

Por otro lado, el tiempo que toma llegar al centro educativo y el medio de transporte utilizado para este fin son influyentes en las predicciones, probablemente, las interacciones, más que cada variable por sí misma sea lo que realmente importa. Por ejemplo, para personas que estén dispuestas a pagar más por un colegio específico, así se encuentre lejos de la vivienda y cuentan con posibilidad de transportarse en ruta escolar, puede que la mayor distancia signifique incluso un mayor gasto en educación. En contraste, una persona que habite en zona rural, cuyo principal medio de transporte sea caballo, mula o una lancha/panchón/canoa, la distancia y la dificultad de arribar al centro educativo a tiempo o en condiciones adecuadas son directamente proporcionales y en consecuencia podrían significar menores incentivos a invertir en educación.

Finalmente, es evidente que la formalidad laboral, las condiciones de vida del hogar y el acceso a tecnología también influyen. De acuerdo con Banerjee y Duflo (2011), las personas más humildes suelen ser más escépticas ante las oportunidades, llevando a ser más complejo vencer el ciclo vicioso de las trampas de pobreza. Hogares humildes que han visto pocas oportunidades educativas a lo largo de su vida y sin acceso a facilidades tecnológicas, de infraestructura o de calidad educativa que los lleve a cambiar su percepción, muy probablemente van a subvalorar los rendimientos de la educación.

Teniendo en cuenta la importancia de las interacciones, que se evidencia en los mejores resultados obtenidos en modelos que capturan no linealidades como los bosques aleatorios, es

necesario reconocer que no solo se requiere focalizar recursos hacia las poblaciones más vulnerables con destinación específica en educación, sino que también se resalta la enorme relevancia de que encuentren un contexto y un entorno propicio para la valoración adecuada de los rendimientos educativos en sus vidas: las facilidades de modo de transporte, acceso a tecnología y condiciones de vivienda, por ejemplo, deben ser también abordadas con urgencia para poder incrementar el gasto en educación de los hogares.

Conclusiones

El gasto en educación es determinante para una economía, siendo el nivel de educación uno de los principales determinantes de ingreso. Si bien los gobiernos provén educación pública a los más pobres, los hogares aún tienen que enfrentar cierto costo asociado a la educación de sus hijos. En los hogares de menor ingreso el gasto en educación es aún más importante, pues la movilidad social va a depender en gran medida del logro educativo de los hijos.

En Colombia, la proporción del ingreso dedicada a la educación es considerablemente baja. La literatura advierte relaciones del gasto en educación con el nivel de ingreso del hogar, el nivel educativo de los padres, el contexto socio económico, el tamaño de la familia, la ubicación geográfica, entre otros aspectos.

En el presente trabajo, se entrenó un conjunto de modelos de distinta complejidad para predecir la proporción del ingreso que destinan a educación los hogares colombianos. Como valor agregado, se penalizó con mayor severidad la sobreestimación, ya que el enfoque está en la identificación de hogares con mayores necesidades de apoyo (gasto más bajo), por lo que la regla de decisión se basó en una métrica propia de los autores.

Es notorio el mejor desempeño de aquellos modelos que capturan no linealidades (como los bosques aleatorios) sobre los modelos lineales (como las regresiones con regularización). Así mismo, la correcta sintonización de hiper parámetros juega un rol clave en la selección del mejor modelo. En este sentido, el análisis de resultados se hizo con respecto al bosque aleatorio que lidera la tabla que se encuentra en anexos.

En consecuencia, se logra predecir con precisión satisfactoria el porcentaje del ingreso que destinan los hogares colombianos al gasto en educación, lo que permite una focalización más precisa de los recursos a hogares con mayores necesidades sacando provecho del aprendizaje de máquinas y el *Big Data.*

Finalmente, se extiende la motivación del documento más allá de la focalización de auxilios económicos, ya que, por medio del análisis de la importancia de variables, así como por las características y resultados de los modelos entrenados, se observa la relevancia de las interacciones de distintos factores del contexto y entorno de los hogares sobre sus decisiones. Por tanto, es clara la necesidad de incidir sobre las facilidades de modo de transporte, acceso a tecnología y condiciones de vivienda, para poder incrementar el gasto en educación de los hogares colombianos. En otras palabras, no solo el ingreso disponible y las características de los miembros de los hogares importan (nivel educativo, aspectos culturales, entre otros), sino que las condiciones del entorno también influyen en lo que los hogares destinan a la educación.

Referencias

Angrist, J., Bettinger, E.m Bloom, E., King, E., & Kremer, M. (2002). Vouchers for Private Schooling in Colombia: Evidence form a Randomized Natural Experiment. The American Economic Review.

Angrist, J., Bettinger, E., & Kremer, M.

Ahmad, Z., & Fatima, A. (2011). Prediction of Household Expenditure on the Basis of Household Characteristics. ResearchGate.

Banerjee, A. y Duflo, E. (2011). Los Mejores de la Clase. En Repensar la Pobreza. (Ed. 4). Penguin Random House GrupoEditorial.

Bayar, A., y Ilhan, B. (2016). Determinants of household education expenditures: Do poor spend less on education? Topics in Middle Eastern and North African Economies, 18.

BID (2021). Impactos del programa Ingreso Solidario frente a la crisis del COVID-19 en Colombia.

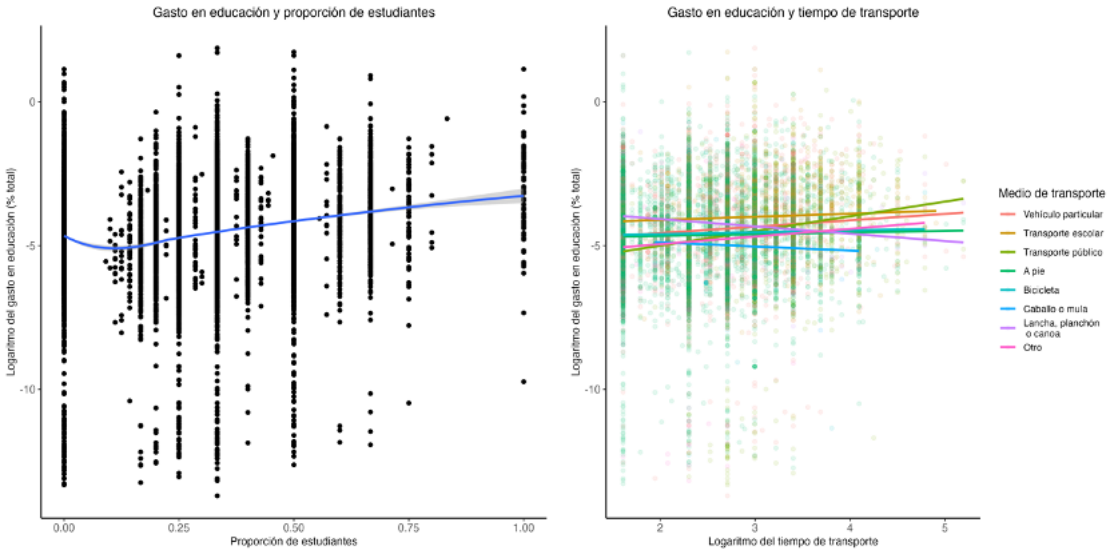
Chi, W., y Qian, X. (2016). Human capital investment in children: An empirical study of household child education expenditure in china, 2007 and 2011. China Economic Review, 37, 52-65.

Deaton, A. (2003). Household surveys, consumption, and the measurement of poverty. Economic Systems Research, 15, 135-159.

Gao, C., Fei, C., McCarl, B., & Leatham, D. (2020). Identifying Vulnerable Households Using Machine Learning. MDPI.

Anexos

Figuras 2a y 2b: Gasto en educación según variables del hogar



Figuras 3a y 3b: Gasto en educación según variables del hogar

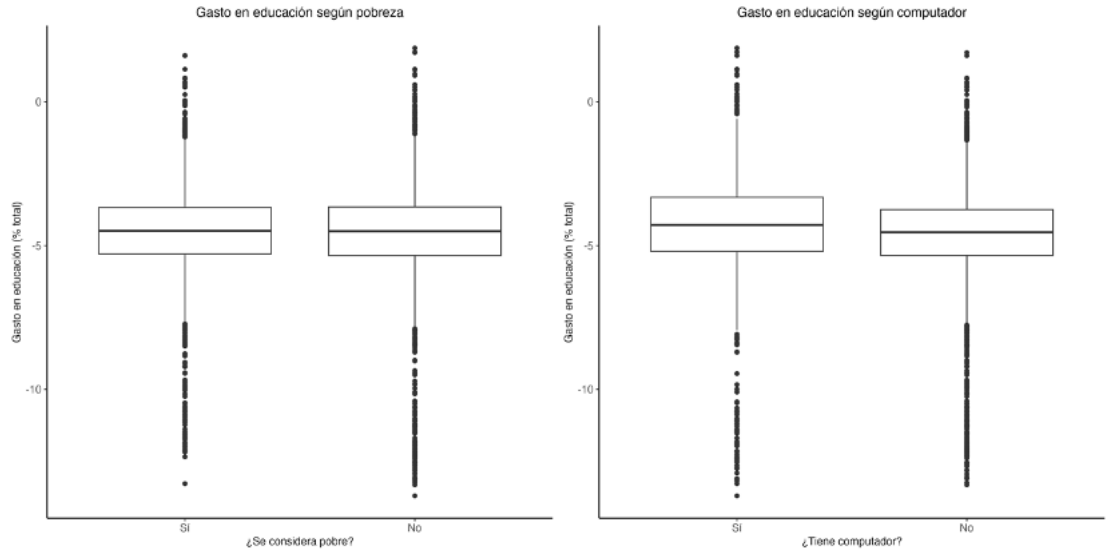


Tabla 2

Modelo	Descripción	RMSE (test)	Personalizada (test)
1	Random Forest de Regresión con 350 árboles; profundidad máxima igual a 20; 40 variables a probar en cada nodo; mínimo numero de observaciones en un nodo para considerar split de 60.Entrenado con RMSE como métrica	0.0340	0.0126
2	Decision Tree Tunning de Regresión con critero de error cuadrático medio; profundidad máxima igual a 7; mínimo numero de observaciones en un nodo para considerar split de 125,mínimo numero de observaciones en un nodo terminal para considerar hoja de 20 . Entrenado con RMSE como métrica	0.0345	0.0126
4	XGBoost con 3000 árboles, tasa de aprendizaje de 0.005; máxima profundidad de 5; gamma de 1; min_child_weight de 100; proporción de columnas por árbol de 0.7; remuestreo con el 60% de los datos. Utilizando la métrica personalizada.	0.0377	0.0128
5	XGBoost con 4000 árboles, tasa de aprendizaje de 0.005; máxima profundidad de 8; gamma de 1; min_child_weight de 100; proporción de columnas por árbol de 0.7; remuestreo con el 60% de los datos. Utilizando la métrica del RMSE.	0.0376	0.0128
3	Bagging Regressor con 600 árboles; 40 variables a probar en cada nodo; sintonización mediante técnica de resampleo OOB . Entrenado con RMSE como métrica	0.0360	0.0135
6	Superlearner que contiene: Un xgboost con 2000 árboles (0.28), un Random Forrest con 500 árboles (0.58), un modelo de elastic net (0.00), una regresión lineal (0.14) y la media de los datos (0.00).	0.0392	0.0136
7	Random Forest de Regresión con 500 árboles; profundidad máxima igual a 10; 15 variables a probar en cada nodo (mtry). Entrenado con la métrica personalizada.	0.0395	0.0137
8	Random Forest de Regresión con 1500 árboles; profundidad máxima igual a 10; 15 variables a probar en cada nodo (mtry). Entrenado con RMSE.	0.0395	0.0138
9	Modelo de Regresión Lineal con Elastic Net. Alpha = 0.0; Lambda = 0.0	0.0505	0.0153
10	Modelo de Regresión Lineal sin Regularización	0.0491	0.0154
11	Modelo de Regresión Lineal con regularización Ridge. Lambda = 0.0254	0.0503	0.0155
12	Modelo de Regresión Lineal con Elastic Net. Alpha = 0.95; Lambda = 0.01	0.0506	0.0158
13	Modelo de Regresión Lineal con regularización Ridge. Lambda = 1.530612	0.0522	0.0166
14	Modelo de Regresión Lineal con regularización Lasso. Lambda = 0.1	0.0525	0.0167
15	Modelo de Regresión Lineal con regularización Lasso. Lambda = 0.1020408	0.0526	0.0167