PROPUESTA DE INVESTIGACIÓN

Predicción del nivel educativo alcanzado para la encuesta "Aprender"

Benitez, Murga, Garcia Oliver

Introducción

La educación es un pilar fundamental en la creación de una sociedad productiva y trabajadora. Es un tema de suma importancia para cualquier sociedad que desee progreso socioeconómico para su nación, y también para cualquier nación que quiera progreso y desarrollo profesional e individual para los ciudadanos que la conforman. En el caso de la Argentina, la educación fue un símbolo de equidad e igualdad de oportunidades para cualquier ciudadano sin importar su clase social, pero las pasadas dos décadas, se vio un claro deterioro en el nivel de la educación, tanto pública como privada, con resultados negativos en exámenes educativos internacionales (evaluación PISA). La destrucción del sector educativo del país se puede deber a varios factores, uno de ellos, para nosotros una de las variables de estudio, es el gasto en educación pública, que estuvo siendo reducido año a año la última década.

Dado los acontecimientos que han sucedido durante los últimos meses referenciando la caída del presupuesto en educación que vienen implementando los últimos gobiernos a cargo del país, se nos ocurrió estudiar cómo impacta sobre el rendimiento académico distintas variables socioeconómicas y demográficas. En esta propuesta, entendemos el rendimiento académico como la tasa de graduación del secundario en todas las escuelas del país. Para analizar y finalmente lograr un modelo predictor de la tasa de graduación, utilizaremos tres bases de datos que procederemos a explicar en el inciso de "Base de Datos", en síntesis nos basaremos en la EPH donde entrenaremos el modelo, en la base de datos "Aprender 2023" donde buscaremos aplicar el modelo entrenado para predecir la variable de interés, y la base de datos de gasto en educación por provincia y jurisdicción que brinda el gobierno nacional.

Intentaremos con estos datos lograr una correcta predicción de la tasa y posibilidad de graduación que tienen los estudiantes que responden las encuestas “Aprender” así poder plantear un ataque en políticas sociales y públicas hacia los estudiantes que se encuentren en posiciones educativas precarias, también poder distinguir el impacto en gasto educativo y entender sí provincias con mayor gasto devuelven estudiantes con mayor nivel educativo. Todo esto será planteado con una metodología incisiva en la comparación de variables y datos de ambas muestras y un importante trabajo de ordenarlas para que las variables estén medidas de la misma manera. Luego planteamos un modelo de árboles(CART) que nos ayudará a mirar nuestros resultados de manera simple y fácilmente presentable a miembros de entidades gubernamentales.

Literatura previa

Las asociaciones entre diversas variables socioeconómicas y el rendimiento académico ha sido explotado por varios investigadores.

Es relevante para nuestro trabajo tener en cuenta la metodología de Valero, S. et. al. (2010). Los autores utilizan los árboles de decisión y vecinos cercanos (KNN) para predecir la deserción escolar analizando datos socioeconómicos y académicos de los estudiantes de la UTIM (Universidad Tecnológica de Izúcar de Matamoros). Por un lado, los árboles de decisión clasifican a los estudiantes en función de atributos como sexo, edad, tipo de bachillerato, promedio de bachillerato, materias reprobadas, apoyos económicos, nivel de inglés, etc. Por otro lado, el algoritmo de KNN utiliza los mismos atributos para identificar estudiantes similares y predecir la deserción en función de sus características más cercanas donde la variable dependiente toma el valor de 1 (Y=1) si el estudiante abandona los estudios y cero (Y=0) si no los abandona. Encuentran que los árboles de decisión y KNN lograron precisiones del 67.07% y 67.77% respectivamente. Las principales causas de deserción son la edad, los ingresos familiares y el nivel de inglés.

Por su parte, Adrogué, C. et. al. (2018) intentan predecir la deserción escolar en la educación secundaria en Argentina mediante un modelo de regresión logística. Este modelo analiza datos de la Encuesta Permanente de Hogares (EPH) para identificar los factores que influyen en la probabilidad de abandono escolar. Las variables analizadas incluyen género, edad, participación en el mercado laboral, tipo de escuela, características del jefe del hogar (género, años de educación y situación laboral), ingreso per cápita familiar, tamaño del hogar, número de hijos y si el hogar es monoparental. La regresión logística se utiliza para estimar la probabilidad de deserción, definida como una variable dicotómica donde 1 indica abandono y 0 indica continuidad en los estudios, en función de estas variables explicativas. Los resultados principales indican que los estudiantes varones, mayores, que trabajan, asisten a escuelas públicas y provienen de familias monoparentales o con menores ingresos tienen una mayor probabilidad de abandonar la escuela. Por el contrario, los estudiantes con mayor ingreso per cápita y aquellos cuyos jefes de hogar tienen más años de educación tienen una menor probabilidad de deserción.

Trigueros (2011) pretende analizar el impacto del gasto público en educación sobre los niveles educativos y los perfiles salariales de los individuos en Colombia, utilizando datos de la Encuesta de Calidad de Vida de 2008. Para lograr esto, se emplean dos enfoques teóricos: la teoría del capital humano, que considera el gasto público en educación como un mecanismo para acumular capital humano, y la teoría de la señalización, que lo ve como una herramienta para que los individuos se señalen en el mercado laboral. Se utilizan métodos como el análisis de regresión y la estimación de ecuaciones estructurales para evaluar estas teorías. Las variables analizadas incluyen años de educación, gasto público en educación por individuo, nivel educativo de los padres y educación promedio del entorno. Los resultados principales indican que el gasto público en educación tiene un impacto positivo pero limitado sobre los años acumulados de educación y los ingresos salariales, y que tanto el enfoque del

capital humano como el de la señalización explican el efecto del gasto público en educación, con un mayor predominio de la teoría de la señalización.

El artículo de Vázquez (2016) es relevante porque analiza la segregación escolar por nivel socioeconómico en Argentina usando datos del PISA. Proporciona información sobre cómo variables socioeconómicas, como ingreso familiar y educación de los padres, influyen en la distribución y rendimiento de los estudiantes. Utiliza diversos índices de segregación para ofrecer una visión completa del fenómeno, ayudando a entender las desigualdades en el rendimiento académico relacionadas con factores socioeconómicos. Además se emplean regresiones lineales . Las variables utilizadas incluyen nivel socioeconómico de los estudiantes, asistencia a escuelas privadas o públicas, y tasas de asistencia bruta al secundario. Los resultados principales indican que la desigualdad y la segregación residencial están positivamente relacionadas con la segregación escolar. Además, la participación del sector privado en la educación incrementa la segregación, especialmente entre los estudiantes del quintil más pobre.

El artículo de Llach y Schumacher (2003) es pertinente para nuestro trabajo ya que proporciona un análisis detallado de la discriminación social en la educación primaria. Utiliza datos del Operativo Nacional de Evaluación de la Calidad (ONE) para identificar cómo el nivel socioeconómico de los estudiantes influye en la calidad de las escuelas a las que asisten y en sus resultados académicos. Los autores emplean métodos como el análisis de regresión por mínimos cuadrados ordinarios (MCO) y la construcción de índices de capital físico, humano y social de las escuelas mediante análisis de componentes principales. Las variables utilizadas incluyen bienes durables de las familias, servicios públicos accesibles, cantidad de hermanos y el nivel educativo de los padres. Los resultados principales muestran que las escuelas a las que asisten los estudiantes de menor nivel socioeconómico son de peor calidad y que esto afecta negativamente sus aprendizajes. Además, se destaca que mejorar la calidad

de las escuelas podría tener un impacto positivo significativo en el rendimiento académico de los estudiantes más pobres. El estudio ofrece propuestas para lograr una mayor equidad educativa a través de la mejora de las condiciones escolares.

El artículo de Serio (2018) es significativo para responder nuestra pregunta de investigación porque tiene como objetivo medir la desigualdad de oportunidades en el rendimiento educativo de los jóvenes en los últimos años de la escuela secundaria en Argentina y analizar la disparidad en las oportunidades educativas para acceder al mercado laboral o a la universidad. Emplea técnicas de análisis de datos complejos, como la estimación de ecuaciones estructurales, para crear índices que reflejan las circunstancias socioeconómicas y geográficas de los estudiantes. Las variables consideradas incluyen la educación de los padres, la ubicación geográfica, los recursos del hogar y el género de los estudiantes. Los resultados principales muestran que, aunque la desigualdad total en los resultados educativos es baja, una parte significativa de esta desigualdad se debe a las circunstancias de los jóvenes, siendo la educación de los padres y la ubicación geográfica los factores más determinantes.

Por último, Chetty et al. (2011) busca evaluar el impacto a largo plazo de los maestros en el rendimiento académico y económico de los estudiantes. Para lograr esto, emplean regresiones lineales y modelos de efectos fijos, utilizando datos del Proyecto STAR, que incluye información detallada sobre los estudiantes y sus maestros, como resultados en exámenes, ingresos futuros y características demográficas. Cabe resaltar que los modelos de efectos fijos son técnicas estadísticas que controlan las variables no observadas que podrían sesgar los resultados, permitiendo aislar el efecto específico de los maestros en los resultados estudiantiles. Los resultados principales indican que los maestros de alta calidad tienen un impacto significativo y positivo en el futuro económico de los estudiantes, aumentando sus

ingresos anuales y la probabilidad de asistir a la universidad. Este estudio demuestra la importancia de la calidad del maestro en la mejora de las oportunidades económicas a largo plazo para los estudiantes.

Metodología y Base de Datos

Para llevar a cabo nuestro trabajo, consideramos necesario la utilización de 3 bases de datos importantes. La primera de ellas es la EPH, esta será la base de datos que usaremos para el entrenamiento de nuestro modelo de predicción, contiene datos de variables de interés para predecir la tasa de graduados del secundario. Esta encuesta contiene datos a nivel individual y a nivel hogar que utilizaremos para relacionarlos con el mismo tipo de datos que encontraremos en la segunda base de datos a utilizar, la encuesta Aprender 2023. La encuesta Aprender 2023, es una encuesta que se hace a estudiantes de escuelas en todo el país. Contiene datos tanto a nivel individual como a nivel hogar, similar a la EPH, esto será necesario para poder aplicar el método predictivo entrenado en la EPH en la base de datos Aprender 2023. La tercera base de datos a utilizar será una que incluye los gastos en educación por jurisdicción/ provincia, esta nos brinda la posibilidad de entender como los gastos en educación impactan en la tasa de graduación del secundario (rendimiento académico). Se deberá combinar esta base con las otras dos, haciendo que cada una de las otras bases incluya los gastos por provincia. Esto lo podremos hacer dado que tanto la EPH como la base Aprender también preguntan por jurisdicción/aglomerado, le podremos asignar a cada persona un nivel de gasto público en su zona.

La metodología que buscamos seguir, es lograr entrenar un modelo con la base de la Encuesta Permanente de Hogares (EPH) para luego poder usar este modelo en la predicción del nivel que un estudiante va a alcanzar de educación dependiendo de ciertas variables incluyendo los gastos en educación por parte del estado. Para lograr esto debemos encontrar

variables similares y comparables en ambas bases de datos, tras un trabajo de análisis encontraremos las variables más relevantes para nuestro modelo.

Variables	**Nombre en Aprender**	**Nombre en EPH**
Edad	ap01	CH06
Sexo	ap03	CH04
Trabaja	ap21b	PP3E_TOT
Escuela Pública o Privada	sector	CH11
Nivel Lectura	ap43	CH09
Lugar de Nacimiento	ap04a	CH15
Aglomerado/Jurisdicción	jurisdicción	AGLOMERADO
Nivel Educativo Alcanzado/ que alcanzara	*Buscamos Predecir*	NIVEL_ED

Tabla 1: Variables Individuales

Variables	**Nombre en Aprender**	**Nombre en EPH**
Nro. Ambientes	ap09	IV2
Nro. de personas que viven en la casa	ap07	Variable creada con la cantidad de personas con el mismo código CODUSU
Parentesco de las personas en convivencia	ap08a, ap08b,ap08c, ap08d, ap08e, ap08f, ap08g, ap08h, ap08i	CH03
Agua potable	ap10a	IV7
Uso exclusivo de un ambiente para estudiar/trabajar	ap10i	II3

Tabla 2: Variables de Hogar

Se deberá hacer un trabajo exhaustivo para lograr hacer comparables las variables similares de ambas bases, así hacer la predicción lo más precisa posible. Por ejemplo, en el caso de “Parentesco de las personas en convivencia” la EPH tiene una variables que contesta a la pregunta, en cambio en Aprender 2023 debemos combinar varias preguntas de la

encuesta en una (ap08a, ap08b,ap08c, ap08d, ap08e, ap08f, ap08g, ap08h, ap08i en 1 que englobe todas). También en el caso de "Aprender 2023" hay datos sobre el nivel educativo de los padres, podríamos, para la EPH, fijarnos en el nivel educativo de los dos adultos más grandes de los hogares para poder crear una variable comparable con la de la base Aprender. Otro paso que podríamos hacer es, usando componentes principales con pesos esparsos para construir un componente que resuma todas las variables de características del hogar en ambas bases de datos, y así comparar esos componentes en vez de cada variable por separado. Esto nos reducirá la complejidad del modelo, haciendo más eficiente el modelo en términos de "Trade-off" sesgo-varianza.

El método que utilizaremos para predecir el nivel educativo al que un estudiante llega es un modelo de árboles de decisión teniendo como referencia el paper de S. Valero, A. Vargas, and M. García (2010). Primero entrenaremos este modelo con la EPH para luego aplicarlo a la base de datos "Aprender 2023" para predecir el nivel educativo que un niño puede alcanzar. Buscaremos poder visualizar qué variables impactan en la tasa de graduados, y que tan significante es el impacto. Este método, como vimos en clase, nos brinda una manera simple y entendible de estudiar el impacto de estas variables (individuales, de hogar y de gasto público) en la tasa de graduación. Permitirá entender cuáles son las características de los alumnos que se reciben del secundario o de aquellos que no. Podremos ver si en jurisdicciones con mayor gasto público en educación consiguen una tasa de graduación mayor, y el gasto en que sector educativo es más importante (educación inicial, primario, secundario).

Para aplicar este modelo debemos encontrar la cantidad de ramas óptima que deba tener el árbol utilizando el método de cross-validación. También se podría utilizar weakest-link pruning para efectivizar el modelo. Entonces, entrenamos el modelo con la EPH y luego lo usamos para predecir en la base de Aprender 2023, y estudiamos los resultados.

Conclusiones y Limitaciones

El modelo que elegimos tiene varias ventajas, una de ellas la simplicidad y la facilidad de entendimiento una vez llevado a cabo. Esto nos facilitará el acto de explicar los resultados a cualquier persona que los necesite. En un tema que afecta a la nación esta persona podría ser una entidad gubernamental que esté interesada en entender los determinantes de la tasa de graduación del secundario y como el gasto gubernamental en educación puede funcionar para aliviar o incentivar la tasa de graduación.

El modelo no tiene solo ventajas, ya que puede ser malo para predecir en el caso de que estemos viendo relaciones lineales, puede ser poco robusto en su predicción. También consideramos que en el caso de nuestra propuesta hay que ser incisivos en lograr ordenar y "Matchear" las variables similares y de interés en ambas de nuestras bases, además de correctamente añadir los valores de gastos a ambas bases de datos. Hay variables que impactan en la tasa de graduación que podríamos estar omitiendo, como el ingreso familiar que no se encuentra en la "Aprender 2023", pero podríamos llegar a entender a la variable "cantidad de habitaciones" como una proxy del ingreso. Teniendo en cuenta los posibles problemas a afrontar y organizando correctamente las bases de datos se podría llevar a cabo una buena predicción de la tasa de graduación de los estudiantes que contestaron la encuesta "Aprender 2023" y determinar en que sectores es mejor invertir en gasto público.

Bibliografía

Adrogué, Cecilia; Orlicki, María Eugenia; Estudiantes en riesgo: un análisis de los factores asociados al abandono de la escuela secundaria en la Argentina desde 2003; Universidad Nacional de Comahue; Pilquen; 15; 1; 7-2018; 21-32

Chetty, Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers, [part] II: teacher value-added and student outcomes in adulthood. The American Economic Review, 104(9), 2633–2679. https://doi.org/10.1257/aer.104.9.2633

Datos de gasto en educación por jurisdicción: https://www.argentina.gob.ar/educacion/evaluacion-e-informacion-educativa/documentacion/gasto-educacion-por-nivel-y-objeto

Edo, M. & Marchionni, M. (2019). The impact of a conditional cash transfer program beyond school attendance in Argentina. Journal of Development Effectiveness.

Llach, J. J., & Schumacher, F. (2004). Escuelas ricas para los pobres. La discriminación social en la educación primaria argentina, sus efectos en los aprendizajes y propuestas para superarla. Publicaciones AAEP. Buenos Aires: Asociación Argentina de Economía Política.

Monserrat Serio (2018). Determinantes de la desigualdad de oportunidades del desempeño educativo del nivel secundario en Argentina.

Riomana Trigueros. (2011). Gasto público en educación: ¿efecto Crowding-in o efecto señalización sobre los niveles educativos y perfiles salariales de los individuos? análisis para Colombia: año 2008. Sociedad y Economía, 1(20), 9–36. https://doi.org/10.25100/sye.v0i20.4056

S. Valero, A. Vargas, and M. García. (2010). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. Recursos Digitales para la Educación y la Cultura, 33-30.

Serio. (2021). Desempeño educativo de los estudiantes en Argentina: Una mirada a la desigualdad de oportunidades del sistema educativo a partir de su medición y descomposición. IDEAS Working Paper Series from RePEc. https://www.argentina.gob.ar/educacion/evaluacion-e-informacion-educativa/base-de-datos

Vázquez. (2012). Segregación Escolar por Nivel Socioeconómico. Midiendo el Fenómeno y Explorando sus Determinantes. IDEAS Working Paper Series from RePEc.