Índice

1 Resumen **2**

2 Justificación **3**

3 Planteamiento del tema **4**

4 Objetivos **4**

4.1 Objetivos específicos: 5

5 Hipótesis **5**

5.1 Subhipótesis 5

6 Marco teórico **6**

6.1 Retornos a la educación 6

6.1.1 Orígenes 6

6.1.2 Estudios empíricos en Argentina 8

6.1.3 Otros estudios empíricos 10

6.1.4 Modelos espaciales 13

6.2 Aprendizaje automático 14

6.2.1 Econometría y aprendizaje automático 14

6.2.2 Convergencia reciente 15

6.2.3 Predicción de desempeño académico 16

7 Metodología y técnicas a utilizar **17**

7.1 Riesgos metodológicos y teóricos 19

7.2 Estructura preliminar de capítulos 20

8 Cronograma **22**

1. Resumen

En el presente trabajo se pretende investigar el poder predictivo, sobre el nivel educativo en hogares, de modelos basados en aprendizaje automático, a partir de microdatos provenientes de la Encuesta Permanente de Hogares (EPH) entre 2003 y 2024. La investigación adopta un enfoque cuantitativo, no experimental, transversal y de carácter predictivo. Se pretende entrenar y evaluar el desempeño de distintos algoritmos de aprendizaje supervisado -bosques aleatorios, SVM y SVR- con selección de variables a través del método de red elástica, para la predicción de indicadores educativos a partir de variables socioeconómicas y demográficas. En base de la generación y evaluación de modelos, se pretende encontrar aquel que permita predecir, con el menor error posible, el nivel educativo de los hogares en Argentina.

El proyecto se concentra en la educación dado que es un determinante clave del bienestar individual y colectivo. En consecuencia, comprender sus causas y distribución es fundamental para disminuir la desigualdad educativa. Para ello, el aprendizaje automático brinda herramientas prometedoras basadas en el hallazgo de patrones complejos.

El objetivo general consiste en desarrollar y evaluar modelos predictivos basados en aprendizaje automático que sean capaces de predecir el nivel educativo en los hogares de Argentina —a partir de la construcción de una métrica homogénea y normalizada—, durante el periodo 2003-2024, sobre la base de variables socioeconómicas y demográficas de la población; con el fin de identificar factores estructurales que expliquen la desigualdad educativa y orienten el diseño de políticas públicas.

El valor agregado del trabajo reside en integrar el enfoque clásico con herramientas contemporáneas, que permiten capturar heterogeneidad y no linealidades, junto con el uso de métricas innovadoras, para lo cual se construyen indicadores del nivel educativo a nivel hogar ajustado por edad y etapa educativa. Esto permite integrar la dinámica familiar dentro de la evaluación.

Estas métricas resultan innovadoras en tanto permiten capturar el déficit educativo estructural en los hogares y, a diferencia de aquellas que se centran en el nivel educativo máximo alcanzado o en el promedio de años de escolaridad, permiten ajustar por edad esperada de escolarización en

cada etapa educativa. Esto posibilita una comparación más equitativa entre hogares con diferentes composiciones etarias e identificar situaciones de déficit educativo de forma más precisa. De este modo, no solo mejora la calidad de la predicción sino la capacidad de focalización de políticas públicas.

2. Justificación

Las estimaciones empíricas sobre las distintas especificaciones de los retornos a la educación, para diferentes periodos y regiones, han demostrado que el nivel educativo es una variable fundamental para el desarrollo económico.

La educación es un determinante estructural del bienestar. El Estado no puede simplemente "asignar" años de educación a aquellos individuos o familias que los necesiten. Por lo tanto, es necesario, para el diseño de políticas públicas, conocer como se constituye el nivel educativo para poder promover cambios significativos.

Un aspecto clave de toda política pública, es poder identificar a sus beneficiarios correctamente para evitar una asignación ineficiente de recursos. El *qué* y *cómo* son tan importantes como el *quién*. En este sentido, contar con modelos eficientes, que permitan identificar a los beneficiarios de una determinada política social, puede ser útil tanto para gobiernos como para ONGs.

La propuesta consiste en, no solo desarrollar modelos de aprendizaje automático que puedan predecir el nivel educativo de los hogares, sino en identificar aquellos factores socioeconómicos y demográficos más determinantes en el nivel educativo. Esto es fundamental para que los hacedores de política puedan identificar las características o atributos de los grupos vulnerables en la sociedad. Por lo tanto, los resultados no son solo predictivos, sino que son interpretables y aplicables como herramientas concretas para mejorar el bienestar agregado.

Considerar el nivel educativo del hogar y no del individuo, permite pensar y analizar las decisiones educativas como respuestas a una dinámica familiar. Esto posibilita poner el foco en los rezagos educativos estructurales en contextos de vulnerabilidad.

3. Planteamiento del tema

El presente proyecto está orientado a investigar el poder predictivo que tienen las variables socioeconómicas y demográficas, cuando se combinan mediante algoritmos de aprendizaje automático (respecto de los métodos econométricos tradicionales), para estimar el nivel de educación de los hogares en Argentina durante el periodo comprendido entre los años 2003 y 2024.

Los resultados obtenidos pueden ser interpretados de forma tal de descomponer las predicciones y entender cuál es el aporte, de cada variable, al resultado final. La identificación de patrones que permitan focalizar adecuadamente las intervenciones, puede contribuir, como insumos para el diseño de políticas públicas, a facilitar las decisiones en cuanto a información y eficiencia.

La disponibilidad de datos ofrecidos públicamente por la EPH y la posibilidad de utilizar herramientas de acceso publico de aprendizaje automático; aseguran la viabilidad técnica y empírica del trabajo.

A partir del planteamiento del tema surgen las siguientes preguntas problematizantes: ¿Es posible predecir con precisión el nivel educativo de un hogar argentino a partir de sus características socioeconómicas y demográficas? ¿Qué algoritmos de aprendizaje automático ofrecen mejor desempeño para esta tarea? ¿Qué variables resultan más relevantes en la predicción del nivel educativo?

4. Objetivos

Objetivo general: desarrollar y evaluar modelos basados en aprendizaje automático que sean capaces de predecir el nivel educativo en los hogares de Argentina —a partir de la construcción de una métrica homogénea y normalizada—, durante el periodo 2003-2024, en función de variables socioeconómicas y demográficas de la población; con el fin de identificar factores estructurales que expliquen la desigualdad educativa y orienten el diseño de políticas públicas.

4.1. Objetivos específicos:

1. Definir y construir una métrica homogénea del nivel educativo ajustada por edad y composición del hogar.

2. Identificar los factores socioeconómicos y demográficos que explican el nivel educativo de los hogares argentinos.

3. Evaluar la capacidad de distintos algoritmos de aprendizaje automático para predecir el nivel educativo en diferentes contextos regionales y temporales.

4. Comparar el poder predictivo de técnicas de aprendizaje automático respecto de modelos econometricos tradicionales en términos de precisión, interpretabilidad y robustez.

5. Interpretar los resultados para proponer estrategias de focalizacion de políticas públicas.

5. Hipótesis

Las características socioeconómicas y demográficas de los hogares argentinos permiten predecir, con alta precisión, su nivel educativo mediante modelos de aprendizaje automático.

5.1. Subhipótesis

El uso de algoritmos de aprendizaje automático, junto con una métrica del nivel de educación por hogar, conlleva a una gran capacidad predictiva del nivel educativo de los hogares en Argentina a partir de variables socioeconómicas observables, lo cual permite comprender nuevas interrelaciones entre las variables.

6. Marco teórico

6.1. Retornos a la educación

6.1.1. Orígenes

La noción de que mayores niveles de bienestar son generados por individuos cada vez más educados, tiene raíces en los orígenes del pensamiento económico formal. Así es como Smith (1776, p. 14) reconocía que "en toda nación, esa proporción [de la riqueza] depende de dos circunstancias distintas: primero, de la habilidad, destreza y juicio con que habitualmente se realiza el trabajo; y segundo, de la proporción entre el número de los que están empleados en un trabajo útil y los que no lo están". En la misma linea, John Stuart Mill argumentaba que el gasto en educación es uno de los pocos que no debería estar vedado para el Estado; Nassau William Senior, vinculaba los avances civiles y educativos con la racionalidad económica; y, Thomas Robert Malthus sostenía que la producción y el mantenimiento de la riqueza, dependen principalmente de las mismas causas, combinadas con la instrucción moral y religiosa (Medema and Samuels, 2004; Malthus, 2024). De forma complementaria, señalaba Marshall (1890, p. 185), que "la habilidad y la destreza industrial dependen cada vez más de amplias facultades como el juicio, la prontitud, el ingenio, el esmero y la constancia —facultades que no están especializadas en un solo oficio, pero que son útiles en muchos— y lo mismo ocurre con la capacidad empresarial" (*traducción propia*).

Sin embargo, el término moderno *capital humano* surge formalmente con Schultz (1961), quien recoge las ideas previas y sostiene que las habilidades y conocimientos adquiridos por los individuos son una forma de capital susceptible de inversión y causante de gran parte del desarrollo humano. Schultz (1961) argumenta que los trabajadores se convertirán, a través de la adquisición de conocimiento y habilidades, en capitalistas. Queda así inaugurada la posibilidad de pensar en la inversión en capital humano como un fenómeno propio del capitalismo.

Becker (1964) profundiza este enfoque y da inicio al desarrollo de una rama de la literatura económica que considera, a la decisión de educarse, como un proyecto de inversión. Para ello, sostiene que el capital humano es análogo al físico y que, por lo tanto, estos pueden analizarse

bajo enfoques similares. Para Becker (1964) es posible elevar el perfil de ingresos futuros de los individuos a través de la inversión en capital humano. Tal inversión tiene asociada así una tasa interna de retorno, que puede ser evaluada bajo los principios del análisis marginal.

A partir de esta conceptualización, Mincer (1974) formaliza la relación entre años de educación, experiencia laboral e ingresos. Su aporte más relevante es la *ecuación de Mincer*, que explica los ingresos como una función que depende de los años de educación y de la experiencia laboral. Esto es,

$$\ln Y_s = \ln Y_0 + rs \tag{1}$$

donde Y_s representa los ingresos hipotéticos que un individuo recibiría luego de completar su escolaridad; Y_0, el ingreso base; r, la tasa marginal de retorno por cada año adicional de educación; y, s, los años de escolaridad. Posteriormente, extiende este modelo para incorporar la inversión post-escolar:

$$Y_{sij} = Y_{si} + \left(r \sum_{t=0}^{j-1} C_{ti} - C_{ji} \right) \tag{1.1}$$

donde Y_{sij} representa el ingreso del individuo i con educación s en el año j de experiencia laboral; C_{ti} y C_{ji}, los costos pasados y actuales de inversión en capital humano; y, r, la tasa de retorno de dicha inversión.

Mincer (1974) halla que, cada año adicional de educación, esta asociado con un incremento de los ingresos entre 7 % y 10 %.

Más allá de las limitaciones del modelo original, este sentó las bases de una serie de desarrollos teóricos y trabajos aplicados, que continúan hasta el día de hoy bajo el paradigma de los modelos mincerianos de retornos a la educación.

6.1.2. Estudios empíricos en Argentina

Los trabajos empíricos sobre retornos a la educación han proliferado a partir de Mincer (1974) en casi todos los países del mundo. Los avances teóricos en econometría y, la disponibilidad de fuentes de información, han permitido el desarrollo de distintas especificaciones de la ecuación de Mincer. Argentina no fue la excepción. Un primer enfoque es presentado por Margot (2001), quien estima los retornos a la educación mediante un análisis basado en cohortes. Para ello, realiza un trabajo aplicado con el objetivo de hallar las tasas de retorno dinámicas para los distintos niveles educativos en Gran Buenos Aires y, compararlas con las tasas de retorno estáticas. Para ello utiliza la siguiente especificación para estimar la tasa interna de retorno de la educación,

$$\sum_{t=0}^{E} \frac{-C_t}{(1+r)^t} + \sum_{t=E+1}^{T} \frac{R_t^j - R_t^{j-1}}{(1+r)^t} = 0 \tag{2}$$

donde C_t representa el costo de oportunidad en el año t; R_t^j, el ingreso acumulado con nivel educativo; j y, R_t^{j-1}, el ingreso con un nivel educativo anterior. La ecuación se iguala a cero para obtener la tasa r que iguala el valor presente de los costos y beneficios educativos.

El principal hallazgo consiste en evidenciar que las tasas de retorno estimadas mediante el modelo estático, son sistemáticamente más altas producto de asumir que los ingresos observados son —sin considerar cambios estructurales ni crisis económicas—, equivalentes entre los distintos grupos etarios. Esto provoca que los ingresos futuros de los individuos resulten sobrestimados y, por lo tanto, las tasas de retorno dinámicas resultan más apropiadas (Margot, 2001).

Estimar los retornos a la educación por mínimos cuadrados ordinarios (MCO), implica asumir que el efecto promedio de un año adicional sobre los ingresos, es homogéneo a través de los diferentes niveles de ingreso. Para ello, Fiszbein et al. (2007) utilizan un enfoque de regresión por cuantiles, para estimar el efecto en los diferentes percentiles de la distribución salarial. A diferencia de MCO, esta técnica permite captar heterogeneidad, en los efectos del capital humano, a lo largo de la distribución del ingreso, bajo la siguiente especificación,

$$\ln W_t = X_i \cdot \beta_\theta + u_{\theta i} \tag{3}$$

donde X_i es el vector de variables exogenas; β_θ, el vector de parámetros; $X_i \cdot \beta_\theta = (Quantile)_\theta(\ln w_i \mid X_i)$, el θ-esimo cuantil condicional del logaritmo del salario, dado X_i, con $0 < \theta < 1$. El θ-esimo cuantil es derivado resolviendo, mediante programación lineal, el siguiente problema: $\text{mín}_{\beta \in \mathbb{R}^k i} \sum \rho_\theta \left(\ln w_i - X_i \beta_\theta\right)$.

Como resultado de estimar la ecuación (3) Fiszbein et al. (2007) hallan que los retornos a la educación se han incrementado año tras año durante el periodo 1992-2002 de 8,6 % a 11,4 %. Estos incrementos, se han mantenido aun bajo periodos de crisis económicas y de deterioro del salario real, acorde a lo expuesto por Schultz (1961). A su vez, observan que en los cuantiles superiores, los hombres tienen retornos mayores respecto del los inferiores. En las mujeres, ocurre lo contrario. En consecuencia, mayores niveles de inversión en educación, conllevarían a mayor desigualdad. Tal desigualdad es provocada por variables en las que los individuos optan voluntariamente, consideradas de esfuerzo personal, como la educación, la migración y ocupación; o de aquellas circunstancias fuera del control individual tales como el género o el lugar de nacimiento (Golley et al., 2019).

La diversidad metodológica en las estimaciones produce resultados diferentes producto de aplicar enfoques diferentes. Cada metodología tiene asociada determinados sesgos. En este sentido, Gómez (2018, p. 12) se propone "analizar empíricamente los retornos a la educación y los premios por calificación en asalariados ocupados de Argentina entre 2003 y 2014, bajo tres especificaciones alternativas de ecuaciones de Mincer". Gómez (2018) considera el ingreso como una función del capital humano, la experiencia y características observables,

$$\ln y_i = \vartheta \left(k_i, e_i, w_i\right) + \ln u_i \tag{4}$$

donde y_i representa los ingresos laborales para el individuo i; k_i, su stock de capital humano medido en años de educación; e_i, su potencial experiencia laboral, aproximada por la edad; w_i, otros

atributos observados del individuo; y u_i, el error aleatorio que captura características inobservables del individuo distribuidas normalmente.

La especificación (4) es estimada por tres métodos: MCO, estimador pseudo-máxima verosimilitud de Poisson y MCO en dos etapas con corrección de sesgo por selección muestral según Heckman (1979). La estimación por MCO ubica los retornos a la educación entre 3,9 % y 5,1 %, con tendencia decreciente desde 2006; Poisson, revela mayores retornos pero también decrecientes; mientras que el método de Heckman exhibe que MCO sobrestima los retornos. A su vez, se observa que las primas por calificación son crecientes entre operarios, pero decrecientes entre técnicos y profesionales (Gómez, 2018).

Gómez (2018) concluye que la metodología de Poisson es la más adecuada para estimar los retornos a la educación porque que permite capturar de mejor manera la evolución temporal y sus estimaciones son más consistentes, incluso bajo la presencia de heterocedasticidad. Respecto del modelo de Heckman (1979), su aplicación se justifica únicamente si existen sospechas de sesgo de selección. Por último, Gómez (2018) destaca que las primas por calificación son menos sensibles al método de estimación, dado que las diferencias entre modelos no resultan estadísticamente significativas.

6.1.3. Otros estudios empíricos

La corrección de sesgo por selección muestral según Heckman (1979) aplicada en Gómez (2018) no es la más adecuada cuando se trabaja con pseudo-paneles. Dado que en estos casos, donde la selección ocurre a nivel de cohortes agregadas, los errores de medición o la falta de variabilidad intra-cohorte, pueden resultar problemáticos.

Para controlar la heterogeneidad no observable entre cohortes y, en consecuencia, generar estimaciones más consistentes, Mora et al. (2023) aplican el método generalizado de momentos corregido (GMMC). Este posibilita capturar efectos fijos de cohortes y, evitar así, los problemas que típicamente se hallan en los datos de corte transversal como la endogeneidad.

Para Colombia, entre los años 2006 y 2020, la tasa promedio de retorno a la educación fue

de 9,7 %. A pesar de ello, durante los últimos años y, a raíz del incremento de la oferta de mano de obra calificada, estas tasas han ido disminuyendo pero conservando una marcada brecha de género (Mora et al., 2023).

Dentro del enfoque de pseudo-paneles, Himaz and Aturupane (2016) estiman los retornos a la educación en Sri Lanka, con un enfoque similar al aplicado por Margot (2001) para Argentina. Sin embargo, mientras Margot (2001) estima la TIR enfocándose en cómo las crisis económicas y el ciclo vital alteran los ingresos proyectados; Himaz and Aturupane (2016) centran su enfoque, mediante uso de efectos fijos, en el impacto que variables no observadas (como la habilidad) tienen sobre la estimación de Mincer.

No controlar por variables inobservables como habilidad y motivación, sesgan la estimación de la igualdad de Mincer por MCO en un 4 % para varones (Himaz and Aturupane, 2016). Similar es la conclusión a la que llega Kemelbayeva (2020) a través de estimar los retornos a la educación en Kazakhstan. Sin embargo, este último explica tal diferencia en base a errores de especificación estructural.

La estrategia adoptada por Kemelbayeva (2020) supone que los retornos a la educación son los mismos a través de los distintos niveles de ingreso. Por otro lado, la regresión por cuantiles, como herramienta, permite desagregar los retornos a la educación por nivel de ingreso utilizando el cuantil como unidad de análisis (Fiszbein et al., 2007). Sin embargo, este método puede estar desatendiendo la existencia de relaciones no lineales. Mahnic (2022) argumenta que los métodos lineales pueden subestimar el impacto de la educación en el crecimiento económico. Se propone "estimar la relación entre los años de estudio de la población y el crecimiento económico mediante el impacto que los primeros tienen sobre el capital humano" (Mahnic, 2022, p. 111). Para ello, parte de la función de producción agregada aumentada por capital humano:

$$Y = AK^{1-\alpha}H^{\alpha} \tag{5}$$

con $0 < \alpha < 1$, donde Y es la producción agregada; K y H, el stock de capital físico y humano respectivamente; y, A, la productividad total de los factores.

Luego de derivar la función, que permite obtener las cotas superior e inferior del capital humano individual, en función de la educación y experiencia laboral; Mahnic (2022) procede a estimar el coeficiente α de la especificación (5). Luego, construye el modelo a estimar,

$$\log y_{it} = \log A + (1-\alpha)\log k_{it} + \alpha \log h^*_{it} + e_{it} \tag{5.1}$$

donde y_{it} representa el nivel de producción per cápita del país i en el momento t; k_{it}, el stock de capital físico; y, h^*_{it}, un factor del capital humano per cápita que depende solo de los años de escolaridad y varia en función de la relación de Mincer considerada.

El autor considera tres formas funcionales para representar la ecuación de Mincer: lineal, cuadrática y cubica. Utiliza datos de 99 países entre 1960 y 2010 para estimar por Mínimos Cuadrados Ordinarios Dinámicos (MCOD) y llega a la conclusión de que la especificación cuadrática ajusta mejor a los datos que la lineal (Mahnic, 2022).

Sin embargo, esto supone considerar los años de educación como unidad homogénea. Para ello Mahnic (2022) controla por calidad educativa según Altinok et al. (2018), para evidenciar de forma más robusta la existencia de no linealidades en la ecuación de Mincer.

En línea con lo propuesto por Mahnic (2022), con el objetivo de capturar efectos no lineales, Kamdjou (2023) propone estimar los retornos a la educación, reemplazando la técnica tradicional de MCO, por un enfoque moderno basado en algoritmos de aprendizaje automático (*machine learning*, ML), en particular regresión de vectores de soporte (*support vector regression*, SVR). Para ello, propone la siguiente función minceriana de ingresos,

$$\log(w_i) = \beta'_0 + \beta_p E_{p_i} + \beta_s E_{s_i} + \beta_t E_{t_i} + \beta'_2 X_i + \beta'_3 X_i^2 + \mu'_i \tag{6}$$

donde w_i representa el ingreso horario del individuo i; E_{p_i}, E_{s_i} y E_{t_i} son variables dicotómicas (dummies) que toman valor 1 si el individuo alcanzó la educación primaria, secundaria o terciaria, respectivamente, y 0 en caso contrario; X_i, la experiencia potencial en el mercado laboral, calculada como la edad menos los años de educación menos 6 ($X_i = \text{edad}_i - S_i - 6$); X_i^2 es el cuadrado

de la experiencia, que permite capturar efectos no lineales sobre los ingresos; β_0', el intercepto de la regresión; β_p, β_s y β_t, los coeficientes estimados asociados a los niveles educativo primario, secundario y terciario, respectivamente; β_2' y β_3', los coeficientes asociados a la experiencia y su cuadrado; y, μ_i', el término de error aleatorio que recoge factores no observados que afectan los ingresos.

Evaluar los retornos a la educación a partir de métodos de ML, permite obtener estimaciones más precisas, robustas y con mejor capacidad predictiva. Su aplicación consiste en entrenar un modelo en base a un conjunto de datos conocidos, con entradas (x) y salidas (y), de modo que este aprenda la función $\hat{y} = f(x)$, a partir de minimizar una determinada función de perdida, entre los valores predichos y los reales. Finalmente, para evitar sobreajuste, el modelo se valida en datos que no han sido usados para su construcción (Kamdjou, 2023).

Producto de estimar (6) mediante SVR, con datos de ocho países representativos de diferentes regiones del mundo, Kamdjou (2023) encuentra que el retorno promedio a la educación se ubica en 10,4 %. Mientras que en América, Asia y Oceanía los retornos se encuentran cercanos al promedio mundial, en África están por encima, en torno al 18 % y, en Europa Occidental por debajo, en torno al 7 %. Kamdjou (2023) concluye que el uso de ML y, particularmente SVR, proporciona estimaciones más precisas y robustas de la ecuación de Mincer que MCO, en parte debido a la inclusión de no linealidades. Sin embargo, los coeficientes estimados mediante ML no son interpretables directamente y, dado que ML está orientado a la eficiencia predictiva, tampoco representan relaciones estructurales ni causales.

6.1.4. Modelos espaciales

Los trabajos empíricos, sobre los retornos a la educación antes citados, muestran que existe una relación positiva entre ingresos y años de educación formal. Los distintos autores hacen uso de diversas aplicaciones. Margot (2001) estima la TIR; Fiszbein et al. (2007) utiliza un enfoque de regresión por cuantiles; Gómez (2018) incorpora los premios por calificación y, los métodos de Poisson y Heckman; y, Himaz and Aturupane (2016) hacen uso de efectos fijos. Todos estos

enfoques explican, de distintos modos, el ingreso a partir de los años de educación como variable explicativa principal. Herrera Gómez (2016) invierte este orden y ubica como variable de interés el nivel educativo en función de la localización geográfica de los centros educativos.

Herrera Gómez (2016) indaga como las características geográficas y sociales influyen en el desempeño educativo. Particularmente se pregunta sobre como la localización de las escuelas impacta en la tasa de repitencia escolar, para el periodo comprendido entre los años 2008 y 2011 con datos de 220 escuelas primarias de la provincia de San Luis.

El modelo construido busca explicar la repitencia, medida como el porcentaje de alumnos repitentes respecto del total, a partir de características propias de la escuela como: su tamaño, el porcentaje de alumnos por profesor, la edad de los profesores, la cantidad de personal no educativo, el porcentaje de alumnos por aula y si la escuela es rural o no; y, de características geográficas a nivel de radio censal, como: el desempleo local, el nivel de hacinamiento, el nivel de analfabetismo y el porcentaje de mujeres en la población (Herrera Gómez, 2016).

Herrera Gómez (2016) estima tres modelos distintos de menor a mayor complejidad incluyendo las variables explicativas progresivamente; para terminar concluyendo que los alumnos con mayor tasa de repitencia tienden a concentrarse en los hogares más pobres y que, en consecuencia, destinar recursos a disminuir la desigualad de oportunidades es equivalente a disminuir la tasa de repitencia y el abandono escolar (Farías et al., 2007). En ocasiones el diseño de políticas educativas está enfocado en los alumnos y no en su entorno. Herrera Gómez (2016) muestra que es tan importante prestar atención al contexto socio-espacial del lugar donde se enseña, como a sus características propias y a las condiciones de la población, para evitar que la movilidad económica se mantenga baja y estable en el tiempo como señalan Neidhöfer et al. (2022) para Latinoamérica.

6.2. Aprendizaje automático

6.2.1. Econometría y aprendizaje automático

Los recientes avances en el poder de computo, sumados a la generación y almacenamiento de grandes volúmenes de datos (*big data*) han impulsado la aparición de un nuevo paradigma (de

análisis empírico) conocido como ML. La generación de modelos de aprendizaje automático, está centrada en la generación de algoritmos flexibles, capaces de capturar patrones completos, con alta capacidad predictiva y sin necesidad de imponer previamente una forma funcional especifica (Breiman, 2001a; Athey and Imbens, 2019; Mullainathan and Spiess, 2017). La econometría surge en una época donde, no solo el poder de computo era menor, sino que la disponibilidad de datos era escasa. Por lo tanto, se desarrollo en línea con la construcción de modelos con pocos predictores seleccionados cuidadosamente en función de la teoría subyacente. Por el contrario, los algoritmos propios del ML son capaces de manejar y sacar provecho del fenómeno de (*big data*) (Charpentier et al., 2018). Este fenómeno, conocido como la *revolución de los datos*, está transformando el análisis económico (Einav and Levin, 2014).

Tanto en econometría como en ML, el enfoque consiste en la construcción de modelos predictivos a partir de un conjunto de datos. Sin embargo, estos difieren en cuanto a metas y enfoques (Mullainathan and Spiess, 2017; Charpentier et al., 2018). Mientras que la econometría se centra en la generación de modelos paramétricos, que permiten explicar fenómenos sociales mediante la estimación de parámetros estructurales, en ML se busca maximizar la precisión predictiva (Mullainathan and Spiess, 2017). Así, Mullainathan and Spiess (2017, p. 88) señala que "el aprendizaje automático gira en torno al problema de la predicción. Lo atractivo es que logra descubrir patrones que se pueden generalizar. De hecho, el éxito de ML en tareas de inteligencia se debe, en gran medida, a su capacidad para identificar estructuras complejas que no fueron especificadas de antemano. Consigue ajustar formas funcionales complejas y muy flexibles a los datos sin caer simplemente en el sobreajuste; encuentra funciones que funcionan bien fuera de la muestra" (*traducción propia*).

6.2.2. Convergencia reciente

Breiman (2001b, p. 109) ilustra que "existen dos culturas en el uso del modelado estadístico para obtener conclusiones a partir de los datos. Una asume que los datos son generados por un modelo estocástico dado. La otra utiliza modelos algorítmicos y trata el mecanismo generador de los datos como desconocido" (*traducción propia*).

En los últimos años, ambas culturas han cooperado en la resolución de problemas vigentes en el análisis estadístico. Un caso de esto, es el problema de selección de variables en contextos en donde la cantidad de predictores p es significativamente mayor a la cantidad de observaciones n. Bajo tales condiciones, estimar por MCO es inestable y, en ocasiones, produce sobreajuste. La técnica *Least Absolute Shrinkage and Selection Operator* (LASSO) permite seleccionar automáticamente las variables más relevantes —algo impracticable mediante métodos tradicionales— con el fin de construir un modelo parsimonioso (Becker, 1964). Utilizar LASSO en contextos de alta dimensionalidad, permite estimar la función objetivo de forma consistente sin perder de vista la relación causal (Mullainathan and Spiess, 2017). Como alternativa al problema de selección de variables se puede aplicar RIDGE que, a diferencia de LASSO, penaliza por L_2 (Athey and Imbens, 2017) o también un enfoque híbrido de red elástica (*elastic net*) que combina ambos tipos de penalización y obtiene las virtudes de cada uno de ellos (Zou and Hastie, 2005).

Otra técnica de ML integrada a la predicción y clasificación de fenómenos de naturaleza económica es la de bosques aleatorios (*random forest*). Esta técnica agregar, a partir de subconjuntos aleatorios de datos, múltiples arboles de decisión. El uso de bosques aleatorios permite reducir el sobreajuste y mejorar la generalización siendo útil en contextos de alta dimensionalidad. Este algoritmo ha demostrado un alto nivel de efectividad en cuanto a predicción y clasificación de problemas económicos (Breiman, 2001b).

6.2.3. Predicción de desempeño académico

Pallathadka et al. (2023) sostiene que mediante el uso de algoritmos de ML, es posible predecir el desempeño académico y, a través de la detección de patrones ocultos, clasificar y orientar las intervenciones correspondientes para evitar o minimizar el fracaso escolar. Para ello hace uso de tres técnicas: naïve bayes y arboles de decisión (*decision trees*) para la predicción y, máquina de vectores de soporte (*support vector machine*, SVM) para clasificación.

Bajo el paradigma de ML, en cuanto a capacidad predictiva, no existe un modelo único dominante, sino que su efectividad depende de la naturaleza de los datos disponibles. Sekeroglu

et al. (2019) en un trabajo similar, utilizan redes neuronales (*neural network*) y SVR para predicción y, gradient boosting para clasificación.

Finalmente, SVM y SVR resultaron ser los mejores algoritmos para clasificar y predecir respectivamente las calificaciones y el rendimiento académico (Pallathadka et al., 2023; Sekeroglu et al., 2019).

7. Metodología y técnicas a utilizar

La presente investigación adopta un enfoque cuantitativo, con un diseño no experimental, transversal, retrospectivo y, de carácter predictivo y comparativo.

La unidad de análisis son los hogares en Argentina, durante el periodo comprendido entre los años 2003 y 2024 inclusive. El INDEC[1], a través de la EPH, recopila información socioeconómica de los hogares y las personas que lo integran.

La variable que se pretende explicar es el nivel educativo del hogar. Para ello, se propone construir una unidad homogénea que permita normalizar y comparar el déficit educativo en los hogares medido como,

$$NE = \frac{\sum IEAE}{N} \tag{7}$$

donde NE es el nivel educativo del hogar; N, la cantidad de miembros mayores a seis años por hogar; $IEAE$ es un indice de educación ajustado por edad, que se calcula como la razón entre los años de escolaridad completada y esperada, donde esta ultima se obtiene restando 5 años a la edad de cada individuo dado que seis años es la edad de ingreso a la escuela primaria. Para individuos de 25 años o más, se propone un máximo de escolaridad esperada de 15 años. Este valor máximo es arbitrario y, conforme se evalúen los datos disponibles se podrá ajustar según la educación promedio de la población.

Una segunda métrica propuesta para aproximar el nivel educativo de un hogar, consiste

[1] Instituto Nacional de Estadística y Censos de la República Argentina

en calcular la proporción de miembros que, teniendo la edad necesaria para haber finalizado una determinada etapa escolar (primario o secundario) no la hayan completado. Se excluye el nivel terciario porque la duración del mismo varia según la orientación elegida.

Las dos propuestas mencionadas, permiten la construcción de una tercera que consiste en agrupar y ordenar en cuantiles los niveles de *ingreso educativo* de los hogares y, así establecer un umbral (por ejemplo, la mediana) como linea de *pobreza educativa* para poder clasificar a los hogares.

A nivel explicativo, las variables que pueden ser útiles para predecir el nivel educativo del hogar son las estructurales del mismo, incluyendo región geográfica, servicios básicos y nivel de hacinamiento; las de composición y organizacional del hogar; y las de ingresos. A su vez, ciertas variables recolectadas a nivel individuo pueden usarse para construir variables agregadas a nivel hogar como genero, ingreso y edad.

Más allá de que la EPH permita utilizar una gran cantidad de variables explicativas, se pretende elaborar un modelo, que sea capaz de predecir eficientemente a partir de únicamente aquellas variables que tengan alto poder predictivo. Esto último puede conseguirse con una técnica de selección de variables como red elástica, siempre y cuando su inclusión sea pertinente y coherente con lo elaborado.

En los casos en los que la variable dependiente sea del tipo ordinal o multinominal se explorará el uso de modelos de clasificación: regresión logística, bosques aleatorios y SVM. Cuando sea continua (como IEAE) se explorará el uso de modelos de regresión: MCO, Bosques aleatorios y SVR. Formalmente, se puede representar como:

$$NE_i = f(X_i) + \varepsilon_i \tag{8}$$

donde NE_i representa el nivel educativo del hogar i; X_i, el vector de variables explicativas (región geográfica, hacinamiento, acceso a servicios básicos, ingreso total, etc.); y, ε_i, el término de error aleatorio que recoge factores no observados.

En cuanto a la clasificación, se puede formalizar como:

$$PE_i = \mathbb{I}(NE_i < \tau) \tag{9}$$

donde PE_i, es una variable binaria que indica si el hogar i tiene déficit educativo; $\mathbb{I}$, es la funcion indicadora que vale 1 si la condicion se cumple y 0 en caso contrario; NE_i, es el nivel educativo del hogar; y, τ, el umbral de pobreza educativa.

La manipulación, purificación y evaluación de los datos se llevará a cabo con el lenguaje de programación Python (versión 3.x), que gracias a su eficacia informática y al extenso ecosistema de librerías enfocadas en el análisis de datos y el aprendizaje automático, posibilitan la realización del presente trabajo (Python Software Foundation, 2023).

7.1. Riesgos metodológicos y teóricos

A continuación se enumeran los principales riesgos metodológicos y teóricos identificados en el desarrollo del proyecto, junto con las estrategias de solución propuestas:

Tipo de riesgo	**Riesgo específico**	**Posible solución**
Metodológico	Algunas variables clave pueden no encontrarse disponibles o estar mal reportadas en la EPH.	Limpieza de datos, imputación múltiple, análisis de sensibilidad.
Metodológico	La definición de años esperados de escolaridad tiene supuestos arbitrarios.	Validación empírica con fuentes externas y pruebas de robustez con diferentes umbrales.
Metodológico	Algunos algoritmos no permiten identificar relaciones interpretables.	Uso de técnicas de interpretabilidad local.
Metodológico	Riesgo de incluir variables redundantes u omitir variables relevantes.	Combinación de métodos automáticos (red elástica) y criterio teórico.
Teórico	Posible tensión entre métodos predictivos y enfoques estructurales.	Enfatizar que el uso de ML complementa, no reemplaza, el análisis económico tradicional.

Cuadro 1: Principales riesgos metodológicos y teóricos

7.2. Estructura preliminar de capítulos

Se propone la siguiente estructura preliminar del trabajo:

1. Introducción

    1.1 Motivación del estudio

    1.2 Preguntas de investigación

    1.3 Estructura del documento

2. Revisión de la literatura

    2.1 Teorías económicas sobre capital humano

    2.2 Retornos a la educación: evidencia internacional y en Argentina

    2.3 Modelos empíricos mincerianos

    2.4 Limitaciones econométricas y alternativas metodológicas

    2.5 Aplicaciones de aprendizaje automático en educación

3. Metodología

    3.1 Diseño metodológico

    3.2 Construcción de la métrica educativa

        3.2.1 Índice de educación ajustado por edad (IEAE)

        3.2.2 Métrica de cumplimiento de etapas

        3.2.3 Umbral de pobreza educativa

    3.3 Variables explicativas consideradas

    3.4 Técnicas de selección de variables

    3.5 Algoritmos de predicción

        3.5.1 Regresión lineal y logística

3.5.2 Random Forest

3.5.3 SVM y SVR

3.6 Evaluación del desempeño predictivo

4. Resultados

4.1 Descripción de la muestra

4.2 Métricas educativas construidas

4.3 Comparación de algoritmos

4.4 Importancia de variables predictoras

4.5 Validación cruzada y errores de predicción

5. Discusión

5.1 Implicancias de los resultados

5.2 Interpretabilidad vs. poder predictivo

5.3 Limitaciones metodológicas

5.4 Comparación con literatura previa

6. Conclusiones y recomendaciones

6.1 Principales hallazgos

6.2 Aportes metodológicos y empíricos

6.3 Recomendaciones para políticas públicas

6.4 Líneas futuras de investigación

8. Cronograma

El siguiente cronograma se presenta como tentativo para la realización del trabajo.

Actividad	**Jun**	**Jul**	**Ago**	**Sep**	**Oct**	**Nov**	**Dic**
Revisión de literatura y antecedentes	X	X					
Construcción de base de datos	X	X					
Definición de métricas educativas		X	X				
Entrenamiento de modelos de ML			X	X			
Evaluación y validación de modelos				X	X		
Interpretación de resultados					X	X	
Redacción de resultados y discusión					X	X	X
Revisión final y presentación del trabajo						X	X

Cuadro 2: Cronograma de actividades: Junio - Diciembre

Referencias

Altinok, N., Angrist, N., and Patrinos, H. (2018). Global data set on education quality (1965–2015). Technical Report Policy Research Working Paper 8314, World Bank Group.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11(1):685–725.

Becker, G. S. (1964). *Human Capital*. NBER, New York.

Breiman, L. (2001a). Random forests. *Machine Learning*, 45:5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Charpentier, A., Flachaire, E., and Ly, A. (2018). Econometrics and machine learning. *Économie et Statistique*, 505(1):147–169.

Einav, L. and Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1):1–24.

Farías, M., Fiol, D., Kit, I., and Melgar, S. (2007). *Todos pueden aprender: Propuestas para superar el fracaso escolar*. Fondo de las Naciones Unidas para la Infancia y Asociación Civil Educación para Todos, Buenos Aires.

Fiszbein, A., Giovagnoli, P. I., and Patrinos, H. A. (2007). Estimating the returns to education in Argentina using quantile regression analysis: 1992–2002. *Económica*, 53:53–72.

Golley, J., Zhou, Y., and Wang, M. (2019). Inequality of opportunity and gender discrimination in China's labour income. In Song, L., Garnaut, R., Fang, C., and Johnston, L., editors, *The*

Chinese Economic Transformation: Views from Young Economists, pages 237–261. ANU Press, Canberra.

Gómez, M. C. (2018). Returns to education and skill premiums: Estimation and biases associated with the case of Argentina. Manuscrito no publicado.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, 47(1):153–161.

Herrera Gómez, M. (2016). ¿La localización de la escuela importa? Condicionantes espacio-contextuales de la tasa de repitencia en un panel de datos georreferenciados. *Documento académico CONICET - IELDE*. Disponible públicamente.

Himaz, R. and Aturupane, H. (2016). Returns to education in Sri Lanka: A pseudo-panel approach. *Education Economics*, 24(3):300–311.

Kamdjou, H. D. T. (2023). Estimating the returns to education using a machine learning approach – evidence for different regions. *Open Education Studies*, 5(1):20220201.

Kemelbayeva, S. (2020). Returns to schooling in Kazakhstan: An update using a pseudo-panel approach. *Eurasian Economic Review*, 10(3):437–487.

Mahnic, P. (2022). Educación y crecimiento económico: considerando no linealidades en la ecuación de Mincer. *Económica*, 68:27–27.

Malthus, T. R. (2024). *Principles of Political Economy Considered with a View to Their Practical Application*. BoD–Books on Demand, Norderstedt.

Margot, D. (2001). Rendimientos de la educación en Argentina: Un análisis dinámico basado en cohortes. In *XXXVI Reunión Anual de la Asociación Argentina de Economía Política*, Buenos Aires. 14 al 16 de noviembre.

Marshall, A. (1890). *Principles of Economics*. Macmillan and Co., London. First edition.

Medema, S. G. and Samuels, W. J. (2004). *The History of Economic Thought: A Reader*. Routledge, London and New York.

Mincer, J. A. (1974). *Schooling, Experience, and Earnings*. National Bureau of Economic Research, New York.

Mora, J. J., Herrera, D. Y., Alvarez, J. F., and Arroyo, J. S. (2023). Returns to human capital in a developing country: A pseudo-panel approach for Colombia. *Economics & Sociology*, 16(1):57–70.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Neidhöfer, G., Ciaschi, M., and Gasparini, L. (2022). Intergenerational mobility of economic well-being in Latin America. Technical Report 303, Documento de Trabajo.

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., and Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. *Materials Today: Proceedings*, 80:3782–3785.

Python Software Foundation (2023). *Python 3.12.0 Documentation*. Python Software Foundation. https://www.python.org/doc/.

Schultz, T. W. (1961). Investment in human capital. *American Economic Review*, 51(1):1–17.

Sekeroglu, B., Dimililer, K., and Tuncal, K. (2019). Student performance prediction and classification using machine learning algorithms. In *Proceedings of the 8th International Conference on Educational and Information Technology*, pages 7–11.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, London. Primera edición.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.