

Análisis explicativo y predictivo de los factores que impactan el rendimiento en las Pruebas Aprender en Argentina

**Autores**

Elina Gandolfo

Belén Garros

Carolina Moni

**Profesores**

Noelia Romero

Ignacio Spiousas

**Asignatura**

Ciencia de datos

**Semestre y año de presentación**

4to año - 2do semestre 2024

**Fecha de entrega**

7/12/2024

Introducción

Las pruebas Aprender son evaluaciones estandarizadas a nivel nacional que se aplican en las escuelas para medir los conocimientos y habilidades de los estudiantes en distintas áreas, como Lengua y Matemática. Estas pruebas están diseñadas por el Ministerio de Educación y suelen aplicarse en escuelas de nivel primario y secundario de todo el país. El objetivo principal de las pruebas Aprender es obtener un diagnóstico del sistema educativo, lo cual ayuda a diseñar políticas y estrategias de mejora. Los resultados también permiten analizar el rendimiento de los estudiantes en distintas regiones del país y en diferentes contextos socioeconómicos.

Esta investigación tiene como objetivo analizar los factores que influyen en el rendimiento de los estudiantes en las pruebas Aprender en la Ciudad Autónoma de Buenos Aires y en la provincia de Buenos Aires, utilizando un enfoque combinado de análisis explicativo y predictivo. La base de datos de estas pruebas ofrece una oportunidad única para comprender las variables que afectan los resultados educativos y para anticipar el desempeño futuro de los estudiantes en función de diversos factores.

Como variables explicativas tenemos características sociodemográficas de los alumnos (edad, género y nivel socioeconómico) y variables escolares (tipo de gestión (pública o privada)), cantidad de días de inasistencia escolar, ubicación geográfica y región de la Argentina). La variable dependiente y a predecir, serían las puntuaciones en Matemática y Lengua. Comprender los factores que influyen en el desempeño en las pruebas Aprender es crucial para el diseño de políticas educativas que apunten a reducir la brecha de rendimiento y a mejorar el sistema educativo en Argentina. Un modelo predictivo basado en estos datos puede ser una herramienta valiosa para anticipar el rendimiento y orientar esfuerzos en áreas críticas.

Como pregunta de investigación surge: ¿Qué factores sociodemográficos y escolares influyen en el desempeño de los estudiantes de CABA y provincia de Buenos Aires en las pruebas Aprender y cómo pueden usarse para predecir los resultados en futuras evaluaciones?

Esta propuesta brinda un aporte interesante a la literatura ya que permite avanzar en la comprensión de los factores que influyen en el rendimiento educativo a nivel provincial, ofreciendo además una herramienta predictiva que puede orientar la toma de decisiones dentro de las políticas educativas. Este análisis puede contribuir al diseño de intervenciones educativas más efectivas y focalizadas, que ayuden a reducir las brechas de rendimiento detectadas entre los estudiantes y a formar un equilibrio dentro del sistema educativo.

Antecedentes

La literatura existente ofrece una base sólida para analizar los factores que influyen en el rendimiento académico, proporcionando elementos clave para esta investigación. Estudios previos, como el de Pérez, Cupani y Ayllón (2005), destacan que el rendimiento académico resulta de una interacción compleja entre habilidades cognitivas, autoeficacia y características de personalidad, siendo las habilidades verbales y matemáticas predictores fundamentales en Lengua y Matemática. Asimismo, los rasgos de personalidad, especialmente la responsabilidad, y el contexto socioeconómico son determinantes relevantes que influyen tanto en el aprendizaje como en los resultados escolares.

El estudio de Tulic (1998), basado en los operativos nacionales de evaluación en Argentina, amplía esta perspectiva al integrar variables escolares y extraescolares. Factores como las expectativas de docentes y familias, el nivel socioeconómico y las prácticas pedagógicas son identificados como determinantes clave en el desempeño. Estos hallazgos destacan la importancia de modelos integrales que expliquen el rendimiento académico, considerando tanto las características individuales como las del entorno educativo.

En esta línea, Adrogué y Orlicki (2022) profundizan en las desigualdades educativas en el marco del operativo Aprender 2017. Sus análisis revelan que variables individuales como género, edad, hábitos de lectura e interés escolar, junto con factores familiares como la educación de los padres y el capital cultural del hogar, influyen significativamente en el rendimiento en Lengua y Matemática. A nivel escolar, aspectos como la gestión privada, el clima disciplinario positivo y el nivel socioeconómico promedio de los estudiantes desempeñan un rol crucial. Además, identifican que la mayor parte de la variabilidad en los resultados (70%) se encuentra dentro de las escuelas, lo que resalta la necesidad de enfoques que aborden tanto las desigualdades estructurales como las características individuales.

Finalmente, estudios internacionales como los de Hanushek y Woessmann (2014), basados en PISA, y análisis regionales como los de Llach y Cornejo (2018), subrayan cómo los efectos de pares y la segregación escolar amplifican las desigualdades en el rendimiento. Estas perspectivas complementan el enfoque local, aportando una visión más amplia sobre los desafíos de la equidad y calidad educativa.

Base de datos

Contamos con una base de datos de las pruebas Aprender de 2023 a nivel alumno, en donde participaron 19.272 escuelas (93,9%) y 614.817 estudiantes (82,2%) tanto del sector estatal como privado y de los ámbitos rural y urbano. Año a año los resultados de la evaluación son publicadas por el gobierno argentino, específicamente la Secretaría de Evaluación e Información (Argentina.gob.ar, 2023).

La base de datos de estudiantes de la prueba Aprender 2023 proporciona una visión detallada del rendimiento académico y el contexto de los estudiantes evaluados. Incluye datos de identificación como los códigos de estudiante, la jurisdicción y la sección escolar, que permiten organizar la información de acuerdo a regiones y localidades específicas. Además, contempla información sobre el sector y el ámbito de las escuelas, destacando si son de gestión pública o privada, y si están ubicadas en áreas urbanas o rurales, lo cual es esencial para entender las condiciones en las que los estudiantes están siendo evaluados.

Los resultados de la prueba están detallados mediante variables que reflejan las respuestas de los estudiantes a preguntas específicas, etiquetadas como “ap01”, “ap02”, etc., y que permiten analizar su rendimiento en distintas áreas del conocimiento. A esto se suman variables de puntaje, como “lpuntaje” y “mpuntaje”, que indican los niveles de desempeño en Lengua y Matemática, respectivamente.

El contexto socioeconómico de los estudiantes también está representado en esta base de datos, con información sobre el nivel educativo de los padres, la migración, la sobreedad y un indicador de nivel socioeconómico estimado (“NSE\_nivel”). Estos datos son fundamentales para analizar la influencia de las circunstancias socioeconómicas y familiares en el desempeño académico. En conjunto, esta base de datos permite realizar un análisis integral sobre el rendimiento estudiantil y los factores asociados en Argentina, facilitando el desarrollo de políticas educativas informadas que busquen mejorar los resultados y reducir brechas de desempeño.

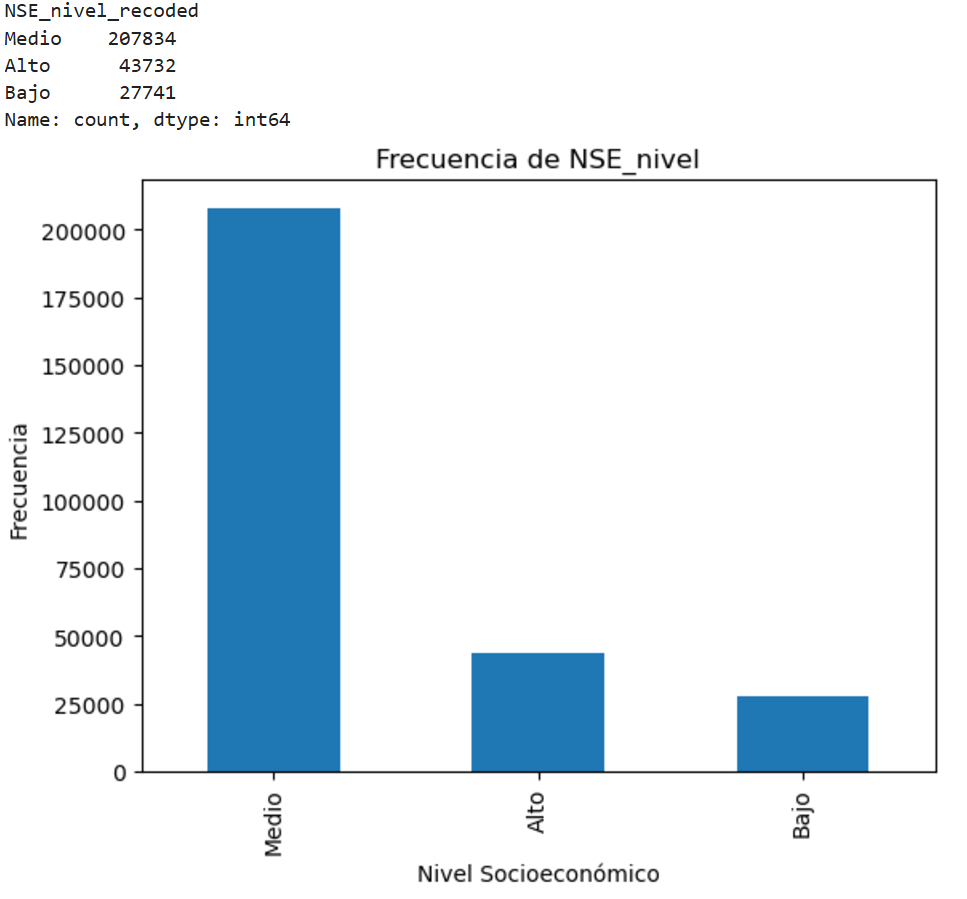
En esta oportunidad analizaremos los datos de las jurisdicciones de la provincia de Buenos Aires y la Ciudad Autónoma de Buenos Aires (CABA), que cuenta con 279.307 datos.

En el análisis de la variable NSE\_nivel (nivel socioeconómico) , que mide el nivel socioeconómico de los estudiantes, se observa que la mayoría de ellos pertenece al nivel Medio, con un total de 207,834 estudiantes. Los niveles Alto y Bajo cuentan con 43,732 y 27,741 estudiantes, respectivamente. No se detectaron valores no asignados en esta variable, lo que asegura una categorización completa y precisa de los datos en este aspecto.

Respecto a la variable ldesemp (desempeño en lengua), que representa el nivel de desempeño en Lengua, los resultados muestran que 98,769 estudiantes se encuentran en el nivel Satisfactorio, mientras que 83,603 están en el nivel Avanzado. Los niveles más bajos, Básico y Por debajo del nivel básico, tienen 51,556 y 27,520 estudiantes, respectivamente. Sin embargo, se identificaron 17,859 valores no asignados. El promedio de desempeño en Lengua, excluyendo estos valores no asignados, es de 2.91, lo que corresponde aproximadamente al nivel Satisfactorio.

En cuanto a la variable mdesemp (desempeño en matemáticas), que mide el nivel de desempeño en Matemática, los resultados revelan que la mayor cantidad de estudiantes (109,633) se ubica en el nivel Satisfactorio, seguido por 61,024 estudiantes en el nivel Básico y 55,544 en el nivel Por debajo del nivel básico. Solo 31,574 estudiantes lograron el nivel Avanzado. También se encontraron 21,532 valores no asignados en esta variable. El promedio de desempeño en Matemática, excluyendo los valores no asignados, es de 2.45, lo que lo sitúa entre los niveles Básico y Satisfactorio.

En general, el análisis muestra que el nivel promedio de desempeño en Matemática es inferior al de Lengua, lo que podría indicar áreas de mejora específicas en el aprendizaje de Matemática. Los valores no asignados en ambas variables, aunque representan un porcentaje relativamente bajo, deberían analizarse para identificar posibles patrones o errores en la recolección de datos.



Frecuencias de ldesemp:

ldesemp\_recoded

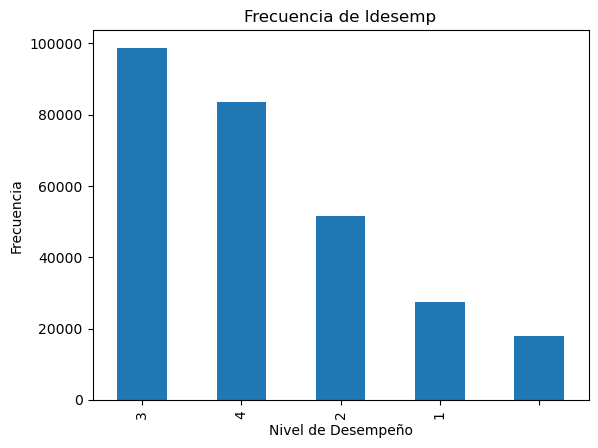
3 98769

4 83603

2 51556

1 27520

17859



Frecuencias de mdesemp:

mdesemp\_recoded

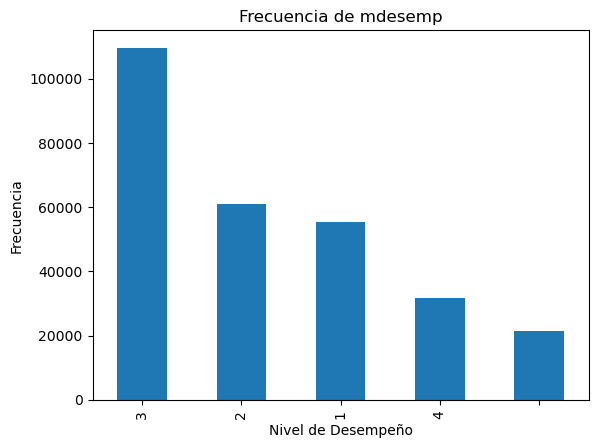
3 109633

2 61024

1 55544

4 31574

21532



Metodología

En este estudio se utilizará una metodología mixta que combina un análisis explicativo y predictivo para ahondar en la comprensión de los factores que inciden en el rendimiento de los estudiantes en las pruebas Aprender. Esta combinación permite identificar las variables clave que afectan el desempeño y evaluar su utilidad para anticipar resultados futuros, brindando tanto conocimiento descriptivo como herramientas prácticas de planificación educativa.

Para identificar las variables que tienen una influencia significativa en los resultados, realizaremos un análisis de regresión que permitirá observar la relación entre características sociodemográficas y escolares de los estudiantes y sus puntajes en Matemática y Lengua. Las variables explicativas incluyen características individuales (género, edad), familiares (educación de los padres, nivel socioeconómico) y escolares (gestión pública/privada, inasistencia). Estas se codificarán en formatos apropiados para el análisis, como variables continuas o categóricas según corresponda.

Para complementar el análisis explicativo, se desarrollará un modelo predictivo. Este modelo será entrenado utilizando los datos de las pruebas de 2016 y validado con los resultados de 2023, permitiendo comparar patrones entre ambos períodos. La evaluación del modelo se realizará mediante métricas como el Error Cuadrático Medio (RMSE) y la precisión, lo que permitirá medir su eficacia y utilidad para futuras aplicaciones.

El análisis será llevado a cabo utilizando software especializado como Python, utilizando librerías como Pandas y Scikit-learn para la gestión de datos y modelado. Además, los datos serán anonimizados y procesados conforme a normativas éticas para garantizar la confidencialidad y protección de la identidad de los estudiantes y las instituciones educativas.

Aunque este enfoque es robusto, una posible limitación es la ausencia de variables no observables, como la motivación estudiantil, que podrían tener un impacto significativo en el rendimiento. A pesar de eso, esta metodología ofrece una perspectiva integral para entender y anticipar las dinámicas del rendimiento académico en el contexto argentino.

Conclusiones y limitaciones

El presente trabajo permitió identificar los principales factores que influyen en el rendimiento académico en las pruebas Aprender en Argentina, destacando el impacto del nivel socioeconómico, la educación de los padres, la gestión escolar y variables individuales como la edad, el género y la asistencia. Los modelos predictivos basados en datos de 2016 y 2023 confirmaron la relevancia de estas variables y evidenciaron la necesidad de estrategias específicas para reducir brechas educativas, especialmente en contextos vulnerables. Estos hallazgos ofrecen una base sólida para el diseño de políticas educativas que mejoren el rendimiento académico y promuevan la equidad en el sistema educativo.

Sin embargo, este estudio presenta ciertas limitaciones que deben considerarse al interpretar los resultados. Aunque la base de datos Aprender es amplia y detallada, no incluye variables relevantes como la motivación o el apoyo emocional, lo cual restringe el alcance del análisis. Además, si bien la comparación entre 2016 y 2023 permitió observar tendencias a lo largo del tiempo, no se pudieron considerar eventos intermedios, como la pandemia de COVID-19, que probablemente alteraron los resultados educativos.

Por otra parte, aunque el modelo predictivo mostró un buen desempeño, su precisión está limitada por la calidad de los datos disponibles y la ausencia de factores no observables que podrían influir significativamente en los resultados.

Finalmente, aunque los hallazgos son representativos del contexto argentino, no necesariamente son generalizables a otros países o regiones con sistemas educativos diferentes. Estas limitaciones abren la puerta a futuras investigaciones que exploren dimensiones adicionales, como aspectos no cognitivos o el impacto de políticas educativas específicas, para ampliar la comprensión sobre los determinantes del desempeño académico.

Bibliografía

Adrogué, C., & Orlicki, M. E. (2022). Factores asociados a los logros académicos en el último año de la escuela secundaria en Argentina. Revista Pilquen, 25(1), 106-129.

Argentina.gob.ar. (2023). *Aprender 2023*. Recuperado de<https://www.argentina.gob.ar/aprender-2023>

Pérez, E., Cupani, M., & Ayllón, S. (2005). Predictores de rendimiento académico en la escuela media: habilidades, autoeficacia y rasgos de personalidad. Avaliaçao Psicológica, 4(1), 1-11.

Tulic, M. L., Cominetti, R., & Ruiz, G. (1998). Algunos factores del rendimiento: las expectativas y el género. LCSHD Paper Series–The Word Bank.