Uso del Aprendizaje Automático para la Predicción del Impacto del Nivel Socioeconómico en las Notas de Lengua de los Alumnos de Sexto Grado en las Pruebas APRENDER 2021

Entrega final

Autoras

Agustina Cirinsky Diez - N° Legajo: 33727

María Emilia Garófalo - N° Legajo: 33726

Luciana Videla - N° Legajo: 34310

Carrera

Licenciatura en Ciencias del Comportamiento

Asignatura

Ciencia de Datos

Profesores

Maria Noelia Romero

Ignacio Spiousas

Tutorial

N° 2 - Hs: 17:20

Fecha de entrega

7 de diciembre de 2024

1. Introducción

Por cuarto año consecutivo, el Ministerio de Educación de la Nación implementó en diciembre de 2021 las pruebas Aprender, donde se evaluó alumnos de sexto grado de la escuela primaria en las áreas de Lengua y Matemática (Ministerio de Educación de la Nación, 2023). Además del rendimiento estudiantil, se incluyeron cuestionarios para obtener información contextual al desempeño escolar mediante el relevamiento de datos vinculados a los atributos individuales, familiares y socioeconómicos, y las trayectorias escolares.

Los resultados del año 2021 indicaron cierta estabilidad en el desempeño de Matemática y una pérdida significativa de aprendizajes en Lengua, que revierte la tendencia a la mejora en los rendimientos iniciada en 2013 y continuada en los sucesivos operativos Aprender. De hecho, un informe de la organización Argentinos por la Educación remarcó que actualmente, el 46% de los estudiantes de primaria no alcanzan el nivel mínimo esperable de desempeño según la prueba regional ERCE en el área de Lengua.

El análisis del Ministro de Educación de la Nación (2023) subraya que el deterioro observado en los desempeños escolares no es homogéneo entre los estudiantes sino que está asociado a factores como el nivel socioeconómico (NSE), la educación de los padres, la escolaridad inicial, la tenencia de libros en el hogar, y el acceso a recursos tecnológicos. Estas disparidades, estrechamente ligadas a desigualdades socioeconómicas, pueden afectar a la asistencia, la participación y el logro de aprendizajes, evidenciando la necesidad de abordar las brechas materiales que sostienen la escolaridad (Ministerio de Educación de la Nación, 2023).

La conclusión del Ministerio es que las políticas educativas deben orientarse a dar una respuesta rápida a estos serios problemas: intensificar la enseñanza de Lengua. No obstante, sería prudente evaluar el impacto real de esta propuesta ya que se podría incrementar la enseñanza y, aún así, no ver resultados significativos en entornos no adecuados.

En este marco, los análisis tradicionales de estadística descriptiva han sido útiles para identificar tendencias generales y correlaciones entre variables clave. Sin embargo, estas técnicas pueden ser insuficientes para capturar patrones más complejos que subyacen en los datos. Bajo la pregunta de investigación *¿De qué manera las combinaciones de factores sociodemográficos y contextuales pueden predecir el bajo rendimiento en Lengua en los modelos de aprendizaje automático supervisados, y cómo podrían utilizarse estos patrones para diseñar intervenciones educativas específicas?*, sería propicio utilizar métodos avanzados de análisis, como técnicas de *machine learning* (ML) para identificar combinaciones de factores sociodemográficos y contextuales que actúen como predictores del bajo rendimiento. Estas técnicas, además, permitirían observar patrones latentes que no son evidentes en el análisis tradicional realizado.

2. Literatura previa

En los últimos años hubo un aumento de investigaciones que aplican técnicas de *ML* en el contexto educativo (Gráfico 1). Forero-Corba (2024) realizó una revisión en 55 artículos publicados en revistas de alto impacto entre los años 2021 y 2023. Treinta y tres de ellos aplicaron técnicas de *ML* e inteligencia artificial (IA), siendo 28 de análisis supervisado. Además, se describieron múltiples aplicaciones que fueron implementadas en contextos educativos con nivel primario, secundario y superior en 38 países. Las conclusiones del metanálisis mostraron el fuerte impacto que tiene el uso de *ML* e IA en este contexto, por ejemplo, detectar el rendimiento académico de los estudiantes de forma temprana.

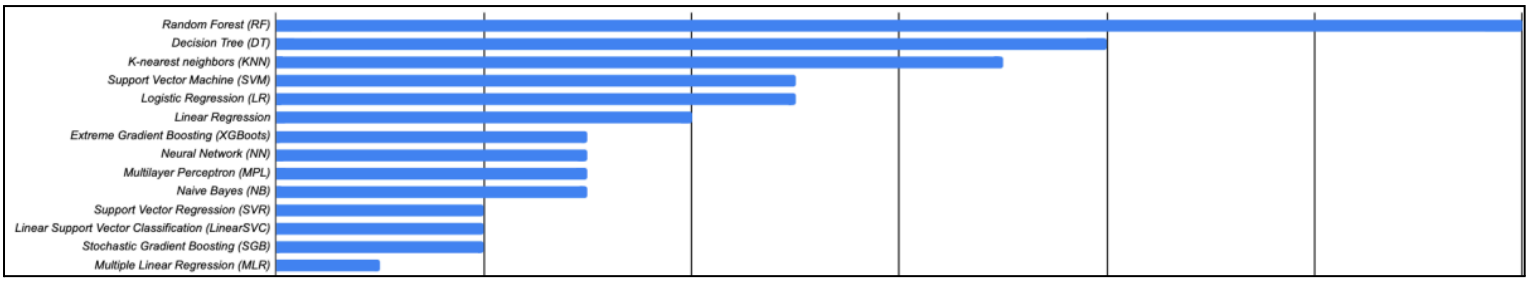

Gráfico 1: Técnicas de aprendizaje supervisado en los estudios revisados (Forero-Corba, 2024)

Si estas técnicas son capaces de generar buenas predicciones de los puntajes, se cuenta con un fuerte indicio de que existen estructuras socioeconómicas que conllevan a distintos desempeños académicos esperados, más aún, se pueden identificar características de dichas estructuras (Vargas, 2022). Es decir que, si bien los programas para evitar el fracaso escolar son importantes, el NSE también tiene un impacto significativo.

En general, las variables que explican este impacto suelen ser el estado civil, la escolaridad, la ocupación y el parentesco del padre, madre o representante; el número de hermanos, la estructura familiar, el ingreso monetario del hogar, los servicios de energía eléctrica, el agua potable regularizado, internet, televisión por cable, entre otros. Estos factores cobran relevancia debido a que los niños comienzan su escolarización en condiciones desiguales para enfrentar las demandas cognitivas y comportamentales que requiere la educación formal y para aprovechar las oportunidades de aprendizaje que ofrece la escuela (Cardozo, 2022). Es decir la "Preparación para la Escuela" (PPE) depende tanto del proceso de maduración y la edad, pero también de la interacción del niño con diversos factores ambientales, familiares e institucionales desde la concepción y durante la infancia.

Cardozo (2022) utilizó información del Panel EIT (Evaluación Infantil Temprana) -un estudio longitudinal sobre las trayectorias educativas de los niños uruguayos -que contiene información sobre el contexto socioeconómico y demográfico de las familias, trayectorias escolares y los fallos de promoción/repetición de los estudiantes de Uruguay; entre otras y realizó un modelo de regresión logística (*logit*). A partir de este determinó que aunque el EIT

fue el predictor más potente, los factores contextuales y socioeconómicos no deben subestimarse, ya que: "sus impactos sobre los resultados escolares[1] [...] actúan en buena medida de forma indirecta, mediados por dichas habilidades" (Cardozo, 2022).

Otro análisis sobre el impacto del NSE en el rendimiento académico fue realizado a partir del uso del modelo CatBoost (Ponce, 2024). Este modelo basado en árboles de decisión se destacó por ser eficaz para manejar desbalances en los datos, que es común en aplicaciones educativas. Para garantizar la precisión del modelo, se optimizaron los hiperparámetros, en especial la tasa de aprendizaje, lo que permitió alcanzar una exactitud del 91% en las predicciones del rendimiento escolar. Además se utilizó SHAP (SHapley Additive exPlanations), que proporciona una explicación detallada de las predicciones para mejorar la interpretabilidad de los resultados. Esto es crucial en el ámbito educativo, ya que permite a los administradores comprender mejor los factores que afectan el rendimiento de los estudiantes, facilitando decisiones informadas basadas en datos.

A través del análisis de los valores SHAP, se identificaron varios elementos socioeconómicos como determinantes del rendimiento académico, que fueron jerarquizados de acuerdo con su impacto. Entre ellos, se destacan la asignatura, las habilidades sociales de los estudiantes, la ocupación de los padres, los ingresos familiares y el número de hermanos. La edad del estudiante, expresada a través del año escolar, y la estructura familiar también se identificaron como variables significativas (Ponce, 2024).

Teniendo en cuenta la búsqueda bibliográfica realizada, se intentará aplicar un modelo de *ML* con alto poder predictivo de las notas de la asignatura de Lengua a partir de tomar como predictores variables socioeconómicas (datos de los alumnos recolectados junto a su evaluación). Es importante determinar cómo estos factores impactan en la nota de la evaluación no solo porque pueden predecir el desempeño académico sino también porque el

[1] Verde: sin dificultades, amarillo: ciertas dificultades y rojo: dificultades severas

aprendizaje de la lectoescritura constituye una de las tareas más importantes de la educación primaria, dado que la habilidad de leer y de escribir incide directamente en el éxito y en el fracaso de niños y adolescentes en la escuela (Urquijo, 2009). En consecuencia, es necesario determinar la forma de estimular la escolarización temprana de los niños de sectores socioeconómicos y culturales más desfavorecidos, con el objeto de compensar los déficits que produce la escasa estimulación para la lectura y la pobreza de sus contextos alfabetizadores (Urquijo, 2009).

3. Base de datos

La base de datos APRENDER contiene todos los resultados de la evaluación Aprender 2021, que revela información sobre el conocimiento de contenidos en Lengua y Matemática, algunos factores socioeconómicos y las condiciones en las que se lleva a cabo la enseñanza, los recursos disponibles y el entorno escolar. La base está disponible para toda persona que quiera consultarla en el portal del Ministerio de Educación de la Nación.

En el año analizado, se implementó la evaluación de forma censal a 19.638 escuelas primarias de Argentina. Un total de 623.558 estudiantes fueron evaluados en Matemática y Lengua, de los cuales 261.696 (41,97%) fueron varones, 271.493 (43.54%) mujeres y 90.369 (14,49%) no binarie. Por otro lado, del total de escuelas, el 61% pertenecieron a zona urbana y el 38,9% a zonas rurales.

Además, se aplicaron cuestionarios complementarios a los estudiantes y a los equipos directivos para obtener información acerca de las condiciones de enseñanza y aprendizaje. Algunas de las características consideradas pueden observarse en los gráficos 2 y 3 y en la tabla 2.

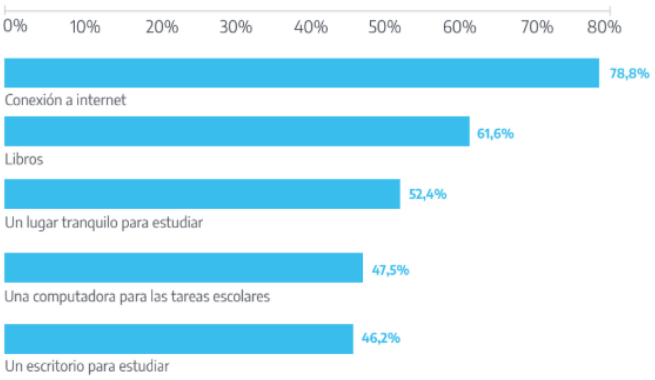

Gráfico 2: Posesión de recursos, equipamiento y servicios en el hogar de las y los estudiantes.

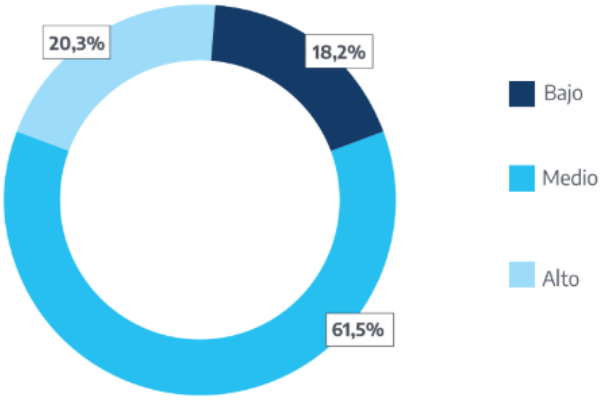

Gráfico 3: Distribución de estudiantes según índice de Nivel Socioeconómico del hogar.

Una de las variables más importantes para nuestro análisis es el puntaje de la evaluación de Lengua. Este no radica en el nivel de habilidad de los estudiantes sino que se interpreta en términos de qué niveles de comprensión de textos logran. El puntaje, entonces, es dividido en cuatro niveles de desempeño (por debajo del básico, básico, satisfactorio y avanzado) de acuerdo a las respuestas de los estudiantes a partir de criterios establecidos por la metodología bookmark realizada por docentes de todo el país en 2016 (Hoszowski y Brenla, 2024). Por otro lado, si bien nuestro análisis se limita a factores específicos acerca del desempeño escolar en relación con el NSE, la base incluye preguntas acerca de la descendencia étnica, el medio de transporte utilizado para llegar a la escuela, factores en relación con las clases virtuales a causa de la pandemia y los motivos de ausencia a clases, la identificación con el grupo de amigos y la comunidad educativa y cuestiones asociadas con bullying,

	Sector de Gestión		Ámbito	
	Estatal	Privada	Rural	Urbano
No fue a la escuela	3,1%	0,5%	4,6%	2,2%
Primario incompleto	6,7%	1,5%	12,3%	4,5%
Primario completo	9,8%	3,3%	17,2%	6,9%
Secundario incompleto	16%	9,5%	15,4%	13,9%
Secundario completo	17,5%	16%	15,5%	17,2%
Terciario/Universitario incompleto	8,5%	9%	6,5%	8,9%
Terciario/Universitario completo	9,5%	19,3%	6,4%	12,9%
Posgrado (especialización, maestría, doctorado, etc.)	6,4%	19,9%	4,1%	11%
No sé	22,5%	21%	18%	22,5%
Total	100%	100%	100%	100%

Tabla 2: Distribución de estudiantes según máximo nivel educativo del padre por sector de gestión y ámbito.

discriminación, maltrato, abusos, derechos, entre otros que serían de gran interés para realizar abordajes varios.

4. Metodología

Considerando los algoritmos de *ML* más comunes utilizados para este tipo de análisis se seleccionarán dos métodos supervisados: *CatBoost* y *logit*. Luego, se compararán sus desempeños para realizar las interpretaciones correspondientes con el mejor de ellos. Respecto de los pasos seguidos, se tomará como referencia los ciclos de vida más populares para estos casos: (1) preparación de los datos, (2) construcción del modelo y (3) evaluación del modelo.

(1) Preparación de los datos: La base de datos "Base estudiantes 6 grado primaria 2021" cuenta con 182 variables que incluyen la nota de las asignaturas e información sociodemográfica. Para comenzar se realizaría una exploración de estos datos, para identificar variables categóricas, numéricas y la variable objetivo. Luego, se deben encontrar aquellas variables categóricas no etiquetadas correctamente (valores como Nan, -6 [multimarca no válido], -9 [blanco]). Más tarde, se revisará el porcentaje de datos faltantes en cada columna y si este es superior al 50%, se eliminará, al igual que los duplicados. Además, podría realizarse un análisis adicional de correlación entre variables antes de decidir eliminarlas ya que quizás se obtendría información útil no evidente en una exploración inicial.

En el caso de haber pocos datos faltantes, si bien *CatBoost* no requiere imputarlos, para *logit* es necesario (media o mediana para variables numéricas, moda para categóricas). En este sentido, *CatBoost* presenta una ventaja, ya que al tener que eliminar filas con variables faltantes (para no introducir sesgos al imputar), *logit* podría quedar con menos datos para predecir. Con el objeto de no perjudicar a *CatBoost*, sería útil separar las bases de datos

y armar una para cada modelo (esto se haría al final cuando ya se hayan tomado todas las decisiones de las variables predictoras). En cuanto a selección de variables, de un total de 182 se eliminarán aproximadamente 80 referidas al ID, al desempeño en matemática y preguntas sobre prevención del abuso sexual y embarazo o discriminación, ya que hay indicadores NSE más relevantes.

Una vez realizada esta selección y habiendo reducido la base a aproximadamente 100 variables, se dividirán las bases de datos para cada modelo. *Catboost* maneja de manera nativa los datos categóricos, por lo que deberíamos convertir estas variables en categorías para evitar que el modelo las trate como numéricas. En el caso de *logit* no se admiten datos categóricos, por lo que habría que aplicar la estrategia de *one-hot encoding*, aunque el problema es que se crearán muchas columnas para cada categoría, por lo tanto, una solución alternativa sería aplicar PCA para reducir la dimensionalidad (que podría comprometer bastante la interpretabilidad del modelo), o crear nuevas variables que se consideren significativas y evaluarlas. Un ejemplo sería combinar variables como: "un escritorio / lugar tranquilo para estudiar", "una computadora que puedas usar para tus tareas escolares", sumando los valores para obtener un puntaje total que refleje el entorno educativo del estudiante y desarrollar un *Índice de recursos educativos* . Así se irían creando otras como *Índice de acceso tecnológico* que abarque las variables "Teléfonos celulares con acceso a Internet", "Computadoras", "Tablets" y "Lector de libros electrónicos". Esta alternativa solucionaría, por una parte el problema de la dimensionalidad y por otra, evitaría una fuga de datos al usar otra codificación como *target encoding* que debe usarse en lugar de *one-hot* debido a la alta dimensionalidad. Otra alternativa a *one-hot* podría ser *frequency encoding* ya que es simple, evita crear muchas columnas y funciona bien para modelos lineales si la relación entre la categoría y el objetivo es evidente, como lo marca la estadística descriptiva.

(2) Construcción del modelo: Para la construcción de los modelos previamente se establecen los datos de entrenamiento (70-80% de los datos) y uno de prueba (30-20%). Se establecerá una semilla para asegurar que en cada ejecución la división de los datos sea siempre la misma, principalmente para que los procesos de *CatBoost* sean consistentes entre ejecuciones. Para la *logit* se agregará la columna de intercepto. Se establecerá *ldesemp*[2] como variable dependiente (categórica: por debajo del nivel básico, básico, satisfactorio, avanzado) y el resto serán las variables independientes. Cabe destacar que se eliminarán todas las relacionadas a las predictoras como lo son el puntaje en números (ya que es la que establece la categoría) y otras que forman parte del puntaje (ej: *lpondera*). Una vez establecido esto, podría realizarse *cross-validation* en los datos de entrenamiento para ajustar los hiperparámetros del *CatBoost* y así evitar un posible sobreajuste, obteniendo una estimación más confiable del rendimiento del modelo.

(3) Luego del testeo de los modelos se evaluará su rendimiento con el *Accuracy* para determinar la proporción de predicciones correctas sobre el total de predicciones realizadas (habría que controlar si hay clases desbalanceadas), una curva ROC para obtener una gráfica que representa la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) a diferentes umbrales de clasificación, y el AUC (Área Bajo la Curva) para medir la capacidad del modelo para clasificar correctamente las instancias de manera separada entre las clases positivas y negativas y asegurarnos que ninguno de los modelos sean peor que una clasificación aleatoria.

5. Conclusiones y Limitaciones

Este trabajo busca encontrar los mejores predictores pertenecientes a factores socioeconómicos para las evaluaciones de Lengua a partir de modelos de *ML* supervisados como lo son *logit* y *CatBoost.* Para ellos se usa la base de datos correspondiente a los

[2] Nivel Lengua

resultados de la prueba Aprender 2021. Si bien con estadística descriptiva tradicional se pueden reconocer variables importantes, como el nivel de escolaridad de los padres, con estos métodos de aprendizaje automático se pueden encontrar interacciones más complejas y factores menos explorados.

Esperamos que las variables más relevantes coincidan con las encontradas en la literatura, pero también es probable que las variables referidas al acceso de la tecnología tengan mayor peso. En 2021 no solo el contexto desfavoreció a familias con menor NSE (alimentación, falta de atención de los padres, trabajo, entre otras) sino que la pandemia restringió el acceso a la escuela, aumentando la dependencia de la virtualidad para el desarrollo de habilidades tanto orales como de lectoescritura.

Este estudio padece algunas limitaciones. Aunque *CatBoost* es más eficiente que *Random Forest* para manejar datos categóricos, sus resultados pueden ser menos interpretables. Además, los índices creados para *logit* hacen que la comparabilidad entre los modelos se reduzca ya que se estarían evaluando otro tipo de variables, agrupadas de manera arbitraria. También podría considerarse el puntaje numérico para capturar mejor la variabilidad entre estudiantes dentro de una misma categoría, lo que permitiría mayor flexibilidad (convertir los puntajes numéricos en categorías después de predecir) y proporcionaría más datos para los modelos.

Por último, estos modelos no solo permitirían un análisis más profundo sino que también facilitarían la simulación de diferentes escenarios para evaluar qué pasaría si ciertas condiciones cambiaran. Por ejemplo: *¿Cómo mejoraría el rendimiento en Lengua si todos los estudiantes rurales tuvieran acceso a recursos tecnológicos equivalentes a los urbanos?* o *¿Qué impacto tendría el aumento de escolaridad de los padres en estudiantes urbanos con bajo rendimiento?* Alentamos a que más investigaciones analicen estos escenarios para evaluar el impacto real de las políticas.

6. Bibliografía

Calderón, V. V., & Ardila, L. F. (2019). Predicción del desempeño en las pruebas Saber 11 utilizando variables del contexto socio-económico de los aplicantes mediante un análisis estadístico con técnicas de machine learning. *Departamento de Física, Universidad Nacional de Colombia*.

Cardozo, S., Silveira, A., & Fonseca, B. (2022). Detección temprana del riesgo escolar. Predicción de trayectorias de rezago en la educación primaria en Uruguay mediante técnicas de machine learning. *Revista latinoamericana de estudios educativos*, *52*(2), 297-326.

Forero-Corba, W., & Bennasar, F. N. (2024). Técnicas y aplicaciones del Machine Learning e Inteligencia Artificial en educación: una revisión sistemática. *RIED-Revista Iberoamericana De Educación a Distancia*, *27*(1).

Hoszowski, A. & Brenla, M.E. (2024). Informe técnico de serie histórica Aprender. Ministerio de Capital Humano.

Tiramontii, G., Nistal, M & Orlicki, E. (2023). Lectura y desigualdad. Comparaciones entre Argentina y América Latina. Observatorio de Argentinos por la Educación. https://argentinosporlaeducacion.org/informe/lectura-y-desigualdad-comparaciones-entre-argentina-y-america-latina/

Ministerio de Educación de la Nación. (2023). *Aprender 2021: Educación Primaria: Las trayectorias escolares de las y los estudiantes de escuelas de nivel primario* (1ª ed.). https://www.argentina.gob.ar/educacion

Pincay-Ponce, J. I., De Giusti, A. E., Sánchez-Andrade, D. A., & Figueroa-Suárez, J. A. (2024). CatBoost: Aprendizaje automático de conjunto para la analítica de los factores socioeconómicos que inciden en el rendimiento escolar. *TE & ET*.

Urquijo, S. (2009). Aprendizaje de la lectura: diferencias entre escuelas de gestión pública y de gestión privada. Revista Evaluar, 9, 19-34.

Urquijo, S., García Coni, A., & Fernandes, D. (2015). Relación entre aprendizaje de la lectura y nivel socioeconómico en niños argentinos. *Avances en Psicología Latinoamericana*, *33*(2), 303-318.