Universidad de
SanAndrés

Big Data

Research Proposal

Profesora: Maria Noelia Romero

Tutora: Victoria Oubiña

Alumnos: Matías Borhi - Manuel Carrera Figueroa

Diciembre, 2023

Introducción

A la hora de interpretar los resultados de una encuesta, el investigador siempre debe tener en consideración el concepto de selección. Es decir, quiénes son aquellos que se han ofrecido a completar el cuestionario en cuestión. Incluso en el caso de las encuestas o evaluaciones obligatorias, uno podría considerar el caso de aquellos individuos que deberían haber respondido la encuesta pues pertenecen al subconjunto de personas para las que la encuesta fue diseñada (por ejemplo, todas las personas empleadas) pero que por alguna razón no han participado en ella. En efecto, la falta de respuesta nos lleva a reflexionar sobre el origen de esa ausencia. Distinguir entre una situación en que la falta de respuesta es más bien contingente de una falta sistemática contiene en sí mismo una gran unidad de valor analítico.

En el mundo de la educación, existe una gran motivación a nivel global de poder entender y conocer la realidad de los alumnos. De ahí que diversas encuestas destinen tiempo y recursos en explorar los resultados académicos de los estudiantes. Por mencionar algunas, a nivel global, las pruebas PISA (Programa para la Evaluación Internacional de los Estudiantes), a escala regional el ERCE (Estudio Regional Comparativo y Explicativo) y en el caso de la Argentina, las pruebas Aprender. Es en este último en el que nos enfocamos en el presente proyecto. Si sometemos la discusión previa acerca de la falta de respuestas al caso de los test educativos, es pertinente cuestionar el alcance verdadero que tienen estas evaluaciones. A priori, encontramos dos subconjuntos de estudiantes para los cuales incluso evaluaciones censales podrían no ser necesariamente representativas. En primer lugar, encontramos el caso de los individuos que precisamente el día en que la evaluación sucedió estaban ausentes. En segundo lugar, están aquellos estudiantes que abandonaron el sistema educativo. Si tenemos en cuenta ambos fenómenos, ausentismo y abandono escolar, las conclusiones obtenidas a partir de las encuestas podrían ser sesgadas. En este trabajo, dada la disponibilidad de datos para Argentina,

trabajaremos utilizando las técnicas de *machine learning* en relación con el ausentismo escolar y las evaluaciones educativas.

El ausentismo es un problema para toda la región latinoamericana y en particular, para el caso argentino (Ministerio de Educación, 2018). A la hora de evaluar el desempeño académico de los niños a través de las pruebas Aprender, la ausencia del alumno se posiciona como un obstáculo directo a la evaluación. Asimismo, al construir proporciones o porcentajes utilizando los resultados de las evaluaciones, los datos faltantes conducen a que los valores no sean representativos de todo el universo de estudio. El problema es solucionado por los analistas encargados de la evaluación a través de un ponderador. Este permite ajustar los resultados de quienes respondieron la prueba para evitar desproporciones y falta de representatividad. Ahora bien, si la ausencia a la prueba fuese aleatoria y estuviese uniformemente distribuida a lo largo de los alumnos, los resultados obtenidos no tendrían mayores niveles de imprecisión. Sin embargo, si la falta a la prueba estuviera correlacionada con algunas variables en particular, las publicaciones del Ministerio de Educación podrían estar sesgadas.

En efecto, en nuestro trabajo, haremos un esfuerzo por mejorar el nivel de precisión de los resultados extraíbles de las pruebas Aprender. Para ello, nos enfocamos en predecir el puntaje académico de aquellos alumnos que, ya sea por ausentismo o alguna otra razón posible, no hubiesen completado la evaluación educativa. A partir de la información presente para los alumnos que sí completaron las evaluaciones, entrenaremos diferentes métodos de *machine learning* para luego hacer las predicciones. Una vez completa esta labor, podremos comparar las distribuciones de puntajes promedio según diferentes categorías (sexo, región, ámbito, sector, nivel socioeconómico) tal como se realiza en los informes formales del Ministerio. En efecto, si encontramos cambios significativos en dichas proporciones, luego nuestro aporte mostraría ser relevante para reducir el sesgo presente en los resultados de las pruebas Aprender.

En primer lugar, la motivación parte del hecho de que el ausentismo escolar es un fenómeno importante dentro del proceso de aprendizaje. Por ejemplo, el BID (2016) asegura que el ausentismo es mayor en escuelas y familias pobres. El hecho de que el ausentismo no sea aleatorio sino que golpee más a ciertos grupos que a otros, nos lleva a querer verificar que las conclusiones de las Aprender sean insesgadas. En segundo lugar, las pruebas Aprender son un instrumento muy útil en el mundo académico, utilizadas en una vasta cantidad de estudios debido a su alcance y nivel de especificidad. Consideramos altamente valioso poder perfeccionar los datos de las bases de datos de Aprender, de modo que las conclusiones que futuras investigaciones obtengan a partir de esta Evaluación, sean lo más precisas posible.

Literatura previa

En cuanto a la literatura previa nos parece destacar un estudio muy reciente llevado a cabo por Acıslı-Celik & Yesilkanat. (2023). En este se utilizan diferentes técnicas de *machine learning* para predecir tanto los puntajes alcanzados en las pruebas de ciencias realizadas en el contexto de las evaluaciones PISA en 2018 como así también los puntajes promedios por países usando como predictores los datos obtenidos en las PISA 2015. En el estudio se utilizaron regresiones lineales múltiples, regresiones *support vector*, *random forest* y *gradient boosting* extremo (*XGboost*). Como resultado de la investigación, se encontró que la metodología *XGboost* es la que mostró el mejor desempeño para la estimación. Este *paper* muestra un ejemplo claro de utilización de diferentes métodos de *machine learning* para realizar predicciones en pruebas relacionadas a la educación como las que planteamos desarrollar en este trabajo.

Otro trabajo que nos gustaría destacar es el de Saarela, Yener, Zaki & Kärkkäinen (2016). En este trabajo se presenta una combinación de modelos de aprendizaje supervisado y no supervisado para la predicción del desempeño de estudiantes en pruebas de matemática, utilizando las bases de datos obtenidas en las pruebas PISA 2012. Se utilizan metodologías

supervisadas para predecir si pudieran pasar las pruebas de matemática de cada nivel. Los modelos que se incluyeron son vecinos cercanos, *Naive* Bayes, análisis de discriminante lineal, *support vector machine* y *random forest*, siendo SVM el que mostró mejor desempeño en la predicción. Este *paper* nos aporta nuevamente el uso de diferentes metodologías para predecir el resultado de pruebas que muestran el nivel de educación basándose en evaluaciones educativas poblacionales. Similarmente, nosotros utilizamos las respuestas de los individuos para los que tenemos información completa para predecir los puntajes de quienes no completaron la encuesta.

Por último, el trabajo de Puah (2021) utiliza metodologías de *machine learning* basadas en regresiones y compara su desempeño con regresiones lineales multidimensionales tradicionales. Esta comparación fue realizada en torno a la predicción de los resultados de las pruebas de ciencia de 2015. Como resultado, se encontró que la precisión predictiva de los modelos no fue sustancialmente diferente aunque sí se encontraron diferencias significativas en los predictores que se identificaron como más importantes en cada caso. Esto nos trae a colación que más allá de que los errores de predicción no sean sustancialmente distintos, tenemos que prestar atención en cómo construye cada modelo dichas predicciones ya que pueden diferir sustancialmente los predictores que toman mayor peso. Es particularmente interesante ya que esta diferencia en los predictores de mayor relevancia fue encontrada en torno a predicciones de resultados de pruebas educativas con la misma naturaleza de las que estamos por evaluar en esta propuesta.

Bases de datos

Para llevar a cabo nuestra investigación, utilizaremos los resultados de las pruebas evaluativas Aprender. Estas evaluaciones tienen como objetivo computar de manera sistemática y abarcativa el desempeño escolar de aquellos alumnos que finalizan su trayectoria educativa

primaria y secundaria. Es por eso que se llevan a cabo una vez al año desde 2016, en ambos ciclos. Las pruebas son estandarizadas y se encuentran segmentadas por área, Matemática y Lengua.

Acerca del alcance de las pruebas Aprender, según el año y el ciclo, la evaluación es censal o muestral. En los casos donde el desarrollo es censal, los resultados obtenidos nos revelan información muy rica sobre el estatus educativo de todos los alumnos del país. Asimismo, las encuestas contienen una gran cantidad de preguntas allende al *outcome* educativo, lo que permite llevar a cabo análisis complejos y desagregados según diferentes variables.

Para iniciar la investigación, nos concentramos en los resultados de las pruebas Aprender de 2022 para secundaria. En esta evaluación, el alcance de los datos es censal, por lo que contamos con una cantidad altamente representativa del alumnado secundario argentino en 2022. A continuación, mostramos brevemente en la Tabla 1 estadísticas descriptivas de las bases a utilizar. Hemos hecho énfasis en aquellas variables que consideramos que podrían tener relevancia a la hora de predecir datos faltantes.

En primer lugar, podemos observar que el puntaje promedio en todos los casos es mayor en lengua que en matemática. También, notamos que la cantidad de datos faltante comprende una proporción importante de la muestra total. Si bien en este caso definimos *missing* como aquellos casos en los que el alumno no muestra puntaje para alguna o ambas disciplinas, en nuestro proyecto, desarrollamos tres versiones, distinguiendo entre falta de puntaje lengua, matemática y ambas.

Asimismo, las mujeres muestran un nivel promedio mayor en lengua, mientras que los hombres enseñan un nivel medio mayor en matemática. El porcentaje de datos faltantes se encuentra balanceado entre ambos sexos. En cuanto al ámbito, los estudiantes del sector urbano muestran un puntaje medio mayor que el de los estudiantes del sector rural para ambas áreas.

TABLA 1: Estadísticas Descriptivas

Variables	*Puntaje promedio y cantidad de observaciones*				*Faltantes*	
Año	Lengua	Total Lengua	Matem.	Total Matem.	Cantidad de *missings*	Ratio
Sexo						
Mujeres	523.13	198,843	467.01	195,970	10,646	5.08
Varones	509.03	176,456	484.10	175,744	9,804	5.28
Ámbito						
Rural	471.80	31,462	450.92	30,944	2,015	6.11
Urbano	518.96	359,829	476.68	355,835	20,580	5.47
Sector						
Estatal	491.35	240,295	458.64	235,930	17,517	6.91
Privado	553.07	150,996	499.61	150,849	5,078	3.26
NSE						
Bajo	469.59	48,858	444.85	47,991	3,566	6.92
Medio	518.90	235,264	474.28	233,233	11,956	4.88
Alto	567.96	56,005	56,12	56,122	1,782	3.01
Sobreedad						
No sobreedad	528.97	297,436	482.46	295,359	13,962	4.51
1 año	471.84	53,496	447.98	52,506	4,285	7.55
2 años	456.88	17,351	440.17	16,962	1,598	8.61
3 años o más	443.66	6,023	436.14	5,858	618	9.54
Nivel Educ. Madre						
Primario	486.58	35,959	453.10	35,420	2,294	6.08
Secundario	508.94	100,176	469.41	99,191	5,389	5.15
Universitario	550.90	112,904	499.40	112,683	4,523	3.86

Fuente: elaboración propia a partir de los resultados de las pruebas Aprender para secundaria 2022. Para la definición de *missings*, determinamos como faltantes aquellos casos en que un alumno no tiene información para alguna de las pruebas (lengua o matemáticas) o para ambas. En el caso del ratio, esta columna muestra el ratio total de *missings* sobre el total de observaciones en la prueba de matemática más el total de *missings* (están expresados en %).

Además, la proporción de datos faltantes es mayor para el caso del ámbito rural. Luego, podemos observar que el sector privado tiene un desempeño medio mayor que los estudiantes del sector público. En esta clasificación, la proporción de datos faltantes en el sector público es más que el doble de dicha proporción para el sector privado. Respecto al nivel socioeconómico, notamos que el promedio de ambas disciplinas es ascendente en el nivel y

que la proporción de faltas es descendente en el nivel. Contrariamente, al mirar sobreedad, la relación es exactamente inversa. Cuanto mayor es la cantidad de años de sobreedad, menor es el desempeño escolar medio pero mayor es la proporción de datos faltantes. Por último, encontramos una relación positiva entre el máximo nivel educativo de la madre y el nivel medio de los estudiantes. No obstante, a menor nivel educativo de la madre mayor es la proporción de datos faltantes.

En suma, las estadísticas descriptivas nos revelan que a priori, parecería ser que la proporción de datos faltantes no está uniformemente distribuida, sino que existen ciertos grupos para los cuales la proporción de datos faltantes es más frecuente. En efecto, consideramos es útil desarrollar nuestra investigación a fin de poder ajustar los datos y proporciones extraídos de las pruebas Aprender.

Metodología

Como señalamos anteriormente, el objetivo principal de este trabajo consiste en predecir los resultados de las pruebas Aprender para aquellos individuos cuyo resultado desconocemos, ya sea en la asignatura de Lengua o Matemática, pero cuyo vector con las restantes características está disponible en la base de datos.

Para ello, el primer paso que debemos realizar consiste en estimar un modelo en base al cuál realizaremos las predicciones para cada uno de los individuos que no cuentan con un resultado para la variable de interés. Sin embargo, para este caso, no entrenaremos un modelo específico que luego sería utilizado para la predicción, sino que emplearemos un conjunto de modelos variados, cada uno con diferentes combinaciones de hiperparametros, de forma tal de elegir aquel que reduzca nuestro error de predicción. Los modelos que serán tenidos en cuenta son: regresión lineal, regresión polinómica, vecinos cercanos, CART, *bagging*, *Random Forest*, *Adaboost* y *Gradient Boosting*. Asimismo, dentro de los modelos en que fuere posible

incluiremos como hiperparámetros los métodos de regularización de Lasso y Ridge con diferentes ponderaciones para estas penalidades, de modo tal que esta variable sea tenida en cuenta al momento de elegir la configuración que minimice el error de predicción.

Antes de profundizar más en la metodología, cabe destacar que será necesario entrenar dos modelos distintos, uno para cada asignatura. Cada uno de estos modelos será utilizado posteriormente para predecir los valores de las respectivas pruebas. En este sentido, además, es necesario mencionar que las observaciones que serán utilizadas para entrenar el modelo en cada caso serán diferentes ya que para cada asignatura, los individuos para los cuales contamos con el resultado de la prueba son potencialmente diferentes.

Retomando la selección de modelos, la configuración de hiperparámetros óptima para cada modelo será elegida a través de un proceso de validación cruzada. Para hacerlo propondremos dividir la muestra inicial en diferentes submuestras. De dicha división un subconjunto de ellas será utilizado para entrenar el modelo bajo análisis y solo una restante para predecir el error por fuera de la muestra. Dicho proceso será repetido alternando las muestras que componen el subconjunto de entrenamiento y aquella que es empleada como testeo. Luego, se calculará el promedio del error cuadrático medio que surgió de las diferentes etapas de entrenamiento y error. Este proceso será repetido para cada configuración sugerida de hiperparámetros, con lo que obtendremos, para cada una de ellas, un error cuadrático medio promedio surgido a partir de la combinación de diferentes submuestras de entrenamiento y testeo.

Una vez que contamos para cada modelo con una lista de configuraciones de hiperparámetros y el error de predicción asociado a dicha configuración, seleccionamos aquella que minimice el error cuadrático medio para cada modelo en particular. Con esto obtendremos una lista de modelos con una configuración de hiperparámetros específica y el error cuadrático medio asociado.

Finalmente, a partir de la comparación de los diferentes modelos, seleccionaremos aquel que minimice el error de predicción. Con aquellos modelos seleccionados por este procedimiento, uno para cada asignatura, realizaremos la predicción por fuera de la muestra de modo de asignar un puntaje para las personas que estuvieron ausentes durante las pruebas pero siguen presentes en el sistema educativo, lo cual nos permite acceder a su conjunto de predictores.

Si bien este modelo con menor error cuadrático medio será aquel que utilizaremos para extraer las conclusiones principales de nuestro trabajo, complementariamente estimaremos los resultados con otros de los modelos y configuraciones sometidos inicialmente a consideración con el objetivo de analizar la robustez de nuestra estimación. Para realizar esta prueba evitando *cherry picking,* pero también cuidando no estudiar la robustez con modelos que se distancian mucho en términos de error de predicción, estimaremos los resultados para aquellos modelos y configuraciones cuya diferencia en el error cuadrático medio no supere el 15% del error cuadrático medio encontrado por el mejor modelo y configuración. Una vez estimados los resultados con estos modelos alternativos, comparemos con el modelo seleccionado inicialmente para asegurarnos que la distancia en las predicciones no sea sustancialmente diferente o en sentidos opuestos. Adicionalmente, para cada uno de estos modelos, estudiaremos los predictores principales con el objetivo de entender si estos modelos con errores cuadráticos medios cercanos toman a las mismas variables como relevantes para estimar la predicción o difieren sustancialmente.

Conclusiones y limitaciones

En conclusión, dada la descripción realizada acerca del fenómeno del ausentismo escolar en Argentina y la estadística descriptiva enseñada, consideramos que los resultados obtenidos podrían arrojar mejoras relevantes en el uso de las pruebas Aprender. Consideramos que los cambios en las proporciones y distribuciones extraídas a partir de las evaluaciones se verían reflejados específicamente en aquellos grupos para los que las faltas son más frecuentes. En

particular, esperamos que los ajustes sean más contundentes, respectivamente, para el caso de los estudiantes que asisten a una escuela pública, en el sector rural, de los niveles socioeconómicos más bajos, para los niños con mayores niveles de sobreedad y cuyos padres tienen menor nivel de educación. Para esos casos en específico, consideramos que las predicciones desarrolladas en nuestra investigación podrían incorporar importantes correcciones a las proporciones y estadísticas de las Aprender.

En cuanto a las limitaciones de nuestro trabajo, en primer lugar, debemos mencionar que nuestras mejoras se lograrían para el subconjunto de estudiantes que no completaron alguna de las pruebas o ambas pero que asisten a la escuela. De modo que, si bien nuestra investigación podría reducir el nivel de sesgo de las estimaciones realizadas a partir de las evaluaciones Aprender, aun así los resultados podrían conservar un cierto nivel de sesgo. Este estaría dado porque las estadísticas acerca del desempeño escolar no consideran a aquellos alumnos que abandonaron el sistema escolar. Dado que la tasa de abandono no es uniforme a lo largo de los distintos sectores, ámbitos, sexos y niveles socioeconómicos (Adrogué, 2018), los resultados obtenidos deberían ser menores para ciertos grupos. Sin embargo, como las pruebas Aprender se desarrollan dentro del grupo de los escolarizados, no podemos solucionar este problema.

En segundo lugar, dada la construcción de las evaluaciones Aprender, no podemos hacer un seguimiento de los alumnos en el tiempo puesto que las pruebas se realizan únicamente en los últimos años de cada ciclo escolar. Podríamos intentar matchear la base de 2016 primaria con la base de 2022 de secundaria, pues los niños que en 2016 finalizaban primaria, en 2022 deberían pertenecer a las bases de secundaria. Sin embargo, dado que se preserva el anonimato de cada niño, solamente podríamos hacer un *match* a nivel escuela.

Referencias

- Adrogué, C., & Orlicki, M. E. (2018). Estudiantes en riesgo: un análisis de los factores asociados al abandono de la escuela secundaria en la Argentina desde 2003.
- Acıslı-Celik, S., Yesilkanat, C.M. Predicting science achievement scores with machine learning algorithms: a case study of OECD PISA 2015–2018 data. Neural Comput & Applic 35, 21201–21228 (2023). https://doi.org/10.1007/s00521-023-08901-6.
- Bos, M. S., Vegas, E., Zoido, P., & Elias, A. (2016). América Latina y el Caribe en PISA 2015:¿ Cómo le fue a la región?.
- Ministerio de Educación, Cultura, Ciencia y Tecnología (2019). Argentina en PISA 2018: Informe de Resultados. Recuperado de https://www.argentina.gob.ar/sites/default/files/argentina_en_pisa_2018_informe_de_resultados.pdf
- Puah, S. (2021, January 28). Predicting Students' Academic Performance: A Comparison between Traditional MLR and Machine Learning Methods with PISA 2015. https://doi.org/10.31234/osf.io/2yshm.
- Saarela M, Yener B, Zaki MJ, Kärkkäinen T (2016) Predicting math performance from raw large-scale educational assessments data: a machine learning approach. In: JMLR workshop and conference proceedings, vol 48, pp 1–8. http://medianetlab.ee.ucla.edu/papers/ICMLWS3.pdf