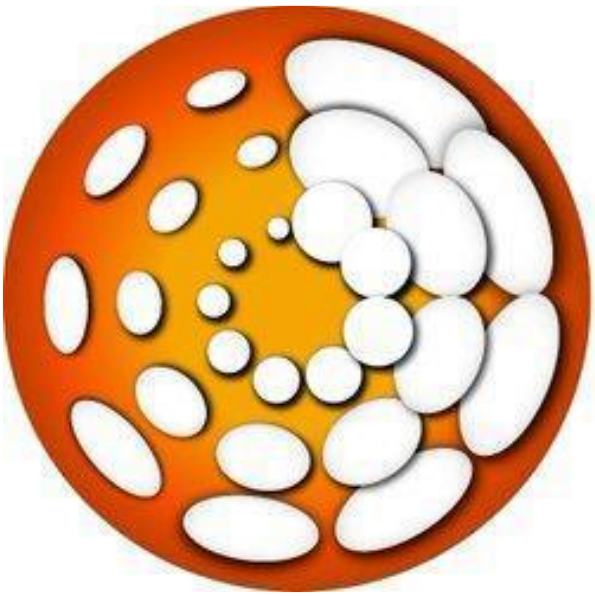
MODELO PREDICTIVO DE PUNTAJES DE LENGUA Y MATEMÁTICA EVALUACIÓN APRENDER 2018



Trabajo Final de Especialización

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Facultad de Ciencias Exactas y Naturales – Facultad de Ingeniería

Universidad de Buenos Aires - Argentina

AUTOR: Lic. Fernando Véliz

ÍNDICE

1. RESUMEN

En el presente trabajo se desarrollaron dos modelos de regresión, a fin de predecir el puntaje obtenido por los estudiantes evaluados en las pruebas Aprender, edición 2018, tanto para matemática como para lengua. A modo de variables predictoras, se tuvo en cuenta el cuestionario sociodemográfico complementario, que además incluye indicios sobre dimensiones relevantes en el proceso de enseñanza y aprendizaje.

Los modelos ajustados muestran un poder predictivo moderado, con una mejoría sobre el nivel basal de alrededor del 20% y un coeficiente de determinación de alrededor del 40%. Estos valores son indicativos de la complejidad del problema, y la dificultad de hacer predicciones con un alto grado de precisión, en base al cuestionario existente. Sin embargo, los modelos muestran como a partir de la administración de un cuestionario complementario limitado, con aproximadamente el 60% de las preguntas originales, es posible predecir con cierto nivel de precisión el puntaje de lengua y matemática esperado, lo cual puede ser útil en situaciones en que por cuestiones de recursos, tiempo o accesibilidad no sea posible administrar el cuestionario complementario completo, junto con las evaluaciones de lengua y matemática propiamente dichas.

Por otro lado, el trabajo revela la importancia de ciertas variables en relación al desempeño académico, como ser el sector de gestión, la existencia de trabajo infantil o el nivel socioecónomico del niño. El desarrollo de políticas públicas que ataquen estas diferencias se presenta como un accionar posible y deseable.

2. INTRODUCCIÓN

En Argentina se implementan evaluaciones nacionales de aprendizajes desde hace más de dos décadas. Aprender es el operativo de evaluación que se desarrolla de forma anual desde el año 2016. Permite obtener información acerca de los niveles de desempeño alcanzados en áreas prioritarias por los estudiantes que se encuentran cursando la educación obligatoria, y sistematizar las percepciones de directivos, docentes y estudiantes, a través de cuestionarios complementarios, sobre dimensiones relevantes en el proceso de enseñanza y aprendizaje.

El objetivo de la evaluación nacional es obtener y generar información oportuna y de calidad para conocer los logros alcanzados y los desafíos pendientes en el sistema educativo y, de esta manera, aportar insumos que contribuyan al diseño de políticas educativas que promuevan procesos de mejora educativa continua y a disminuir las brechas de inequidad existentes.

La evaluación Aprender, siguiendo la línea de trabajo iniciada con el Operativo Nacional de Evaluación (ONE), es una prueba referida a criterios. Estas pruebas buscan conocer los contenidos y capacidades que los estudiantes dominan, a través de un conjunto de ítems relevantes y representativos de la disciplina evaluada. En particular, se evaluaron a estudiantes de 6° grado de nivel primario, en dos temáticas: lengua y matemática. El operativo pretende representar a la totalidad de la población infantil de Argentina cursando el 6° grado.

El presente trabajo se propone realizar dos modelos de regresión, para predecir el puntaje obtenido por los estudiantes evaluados en las pruebas Aprender, edición 2018, tanto para matemática como para lengua. Estos modelos pueden ser útiles a fin de encontrar variables más correlacionadas con el rendimiento escolar en lengua o matemática, a fin de focalizar los esfuerzos del Estado en la mejora de dichas variables, logrando de esta forma alcanzar de forma más equitativa al conjunto de la población escolar. Estos modelos podrían además ser utilizados a fin de tener una estimación a priori, dentro de cierto nivel de confianza, de los resultados educativos esperados en determinada escuela o comunidad, ya sea a partir de contar con información demográfica, o bien administrando solamente el cuestionario complementario. De esta forma, se podría contar con reportes con mayor periodicidad a la anual, o bien estimar los puntajes esperados en aquellos casos en que no se pudiera aplicar la prueba.

2.1. SOBRE LAS PRUEBAS APRENDER

Según nos indica el documento metodológico del Ministerio de Educación [8]:

Aprender es una prueba referida a criterio, por lo que tiene un marco de referencia, esto es, un conjunto delimitado de conocimientos y habilidades evaluados. En el caso de Aprender, ese conjunto consiste en los Núcleos de Aprendizaje Prioritarios, los diseños curriculares jurisdiccionales y los consensos construidos federalmente. Los ítems de las pruebas correspondientes a Aprender se analizaron con base en la TRI (Teoría de Respuesta al Ítem), modelo general sobre el cual se basan la mayoría de las evaluaciones estandarizadas internacionales, así como también los ONE desde 2005.
Aprender administra al conjunto de alumnos 72 ítems diferentes, en cada año y disciplina. Como a un alumno de nivel primario o secundario le resultaría dificultoso responder esta cantidad de ítems, se construyen 6 modelos de cuestionarios, cada uno con 24 ítems. Algunos de estos modelos comparten una parte de los ítems.
Se complementa esto con un cuestionario complementario que enmarcan los desempeños en sus contextos, aplicados a los estudiantes participantes y a los directores de las escuelas participantes.

Interpretación de los puntajes TRI

La escala TRI es arbitraria (tanto su valor medio y desvío standard). No hay un cero natural, como puede ser cuando se puntúa a un alumno contando la cantidad de respuestas correctas. Por ejemplo en ONE 2013 el valor medio de la escala de referencia fue 0 y su varianza 1. En PISA o Aprender, la escala de referencia tiene media 500 y desvío standard 100.
Los puntajes TRI por sí mismos no miden lo que los alumnos saben, tampoco sus niveles de habilidad. La interpretación de los puntajes en términos de qué niveles de problemas los niños saben resolver o qué niveles de interpretación de textos logran, se hace mediante la definición de niveles de desempeño, cuatro en las evaluaciones Aprender. Hay una diversidad de metodologías para determinar los niveles de desempeño y los puntos de corte en la escala TRI (o en otras escalas) que los definen. En Aprender se aplicó el Método Bookmark, también aplicado en otros Institutos de Evaluación Educativa de América Latina. Para esta tarea, fueron convocados un conjunto de docentes con tarea frente a aula, seleccionados en forma aleatoria y representativos de las distintas jurisdicciones y sectores de gestión, quienes en los Talleres Bookmark, luego de tres jornadas de deliberaciones, definieron los puntos de corte y la descripción de cada uno de los niveles de desempeño. Este trabajo se hizo para cada disciplina y cada año evaluado.

En las bases se presenta solo el puntaje obtenido por el alumno y el nivel de desempeño correspondiente (por debajo del básico, básico, satisfactorio y avanzado).

Comparación con otras evaluaciones

Los niveles de desempeño determinados por los especialistas están asociados a una prueba en particular, que evalúa la capacidad de resolver ciertos problemas acordes a ciertos saberes que se supone deben manejar aproximadamente los alumnos de ese año y disciplina. En el caso de Argentina, se toma como referencia los Núcleos de Aprendizaje Prioritario. La categorización de los diferentes puntajes obtenidos en alguno de los niveles, y en particular en el nivel Satisfactorio, lo hacen los docentes convocados a los Talleres Bookmark en base a su experiencia de aula. Por esto, no es posible establecer una comparación directa con otras evaluaciones, que toman como referencia otros conocimientos o habilidades supuestas conocidas por alumnos de la misma edad y con niveles de desempeño definidos por docentes con experiencias de aula, sistemas educativos y sociedades distintas.

La comparabilidad en el tiempo de las pruebas Aprender

Aprender 2016 se fijó como escala de referencia. En este año, la media en todas las disciplinas evaluadas fue 500, y su desvío estándar 100. Para poder comparar los resultados de las pruebas a lo largo del tiempo y que un aumento o disminución del puntaje refleje un aumento o disminución en las competencias y no una prueba en promedio más difícil o más fácil, las pruebas comparten en las sucesivas evaluaciones, para cada disciplina y año, un conjunto de ítems en común, el bloque de anclaje.

2.2. PONDERACIÓN

Las bases cuentan con ponderadores para los niveles de desempeño y para los cuestionarios complementarios. Tal como es indicado en el documento metodológico [8], todos los cruces de variables que involucren a los puntajes utilizan los correspondientes ponderadores (“lpondera” para lengua y “mpondera” para matemática). Para los análisis exploratorios relacionados con datos demográficos o demás variables que surjan del cuestionario complementario, y no involucren puntajes, se tomó en cuenta el ponderador “pondera”.

La ponderación es una práctica usual en encuestas sociodemográficas, que permite representar a la totalidad de una población, aún cuando no se hayan llegado a censar la totalidad de los individuos. Si bien todas las instituciones y estudiantes fueron convocados a estas pruebas, por diversas razones hay escuelas o estudiantes que no pudieron participar. Es por eso que hay áreas evaluadas de forma muestral, seleccionadas de forma probabilística, de tal forma que sean representativas de los distintos niveles, como las jurisdicciones, sector de gestión, ámbito, etc.. Es en estos casos que cobra relevancia la función del ponderador.

3. METOLODOLOGÍA

3.1. LIMPIEZA DE DATOS

La base inicial puesta a disposición por el Ministerio de Educación [9] nos muestra un total de 585.292 registros. Esto representa una pequeña diferencia con el total de estudiantes relevados según el informe: 573.939. A continuación, eliminamos los registros:

1) Con valores nulos en los puntajes lengua o matemática (40.215).
2) En una segunda instancia, con valores nulos en el cuestionario demográfico (4.389).

Esto nos deja un total de **540.688** registros válidos.

3.2. VARIABLES DESCARTADAS

No se utilizaron como variables predictoras las siguientes variables presentes en el conjunto de datos:

- 'isociol_puntaje','isociom_puntaje', ‘ ‘isocioal’, ‘isocioam’: no encontramos justificación teórica para estas variables. Por otro lado, guardan una fuerte similitud con ‘isocioal_puntaje’ y ‘isocioa’.
- ‘dificultad_lengua’, ‘dificultad_matematica’, ‘dificultad_cuestionario’, ‘dificultad_vision’: suponen la administración previa de los cuestionarios, lo que es invalidado por el objetivo del modelo.
- ‘jardín’, ‘repitencia’, ‘compañeros’: por tratarse de variables repetidas

3.3. TRATAMIENTO DE VALORES NULOS

Unificamos los siguientes valores con la codificación -9 (nulos):

-1. No corresponde

-6. Multimarca

-9. Blanco

El porcentaje de valores nulos es muy oscilante según la variable. Variables como edad, país o sexo tienen solamente alrededor de un 1% de valores nulos, mientras todo el conjunto de variables asociadas a la parte del cuestionario realizada a estudiantes de escuelas rurales tiene más de un 90% de valores nulos (en tanto el ámbito rural representa el 11% de los casos). Es por esto que decidimos **no utilizar** para los modelos predictivos todas las variables asociadas exclusivamente a estudiantes de **escuelas rurales**.

Se utilizaron tres estrategias a la hora de imputar valores nulos:

1. Las variables categóricas abiertas en variables *dummies* de tipo existe/no existe (por ej., vive con padre, vive con madre, vive con tío, etc.), o bien las variables dicotómicas (por ej., es hogar migrante, o pertenece a hogar indígena) fueron imputadas con 0, asumiendo que la no respuesta implica inexistencia de la característica. Apelamos a este supuesto para no tener que incrementar innecesariamente el número de variables en la base.
2. Las variables estrictamente categóricas no fueron imputadas, en tanto el valor nulo puede considerarse como una categoría más.
3. La variable 'isocioa_puntaje'. Esta variable está estandarizada en su origen, por lo que consideramos pertinente la imputación de valores nulos con 0, es decir, la media.

3.4. TRANSFORMACIÓN DE VARIABLES CATEGÓRICAS

Las variables con valores binarios fueron dejadas tal como están. Transformamos las variables categóricas en formato columnar, con una nueva columna por categoría, a fin de que puede ser interpretada por los diferentes algoritmos utilizados. Por otra parte, los valores nulos son codificados dejando todas las columnas en 0.

Por ejemplo la variable “Aproximadamente, ¿cuántos libros hay donde vivís?” consta de 5 categorías (los casos con categoría “No sé” se convirtieron a nulos):
1. No hay libros
2. De 1 a 25 libros
3. De 26 a 50 libros
4. De 51 a 100 libros
5. Más de 100 libros

En la transformación implementada, obtenemos la siguiente codificación:

Valor original	cant_libros_viv_1	cant_libros_viv_2	cant_libros_viv_3	cant_libros_viv_4	cant_libros_viv_5
1. No hay libros	1	0	0	0	0
2. De 1 a 25 libros	0	1	0	0	0
3. De 26 a 50 libros	0	0	1	0	0
4. De 51 a 100 libros	0	0	0	1	0
5. Más de 100 libros	0	0	0	0	1
Nulos	0	0	0	0	0

4. ANÁLISIS EXPLORATORIO

4.1. DISTRIBUCIÓN DE VARIABLES DESTACADAS

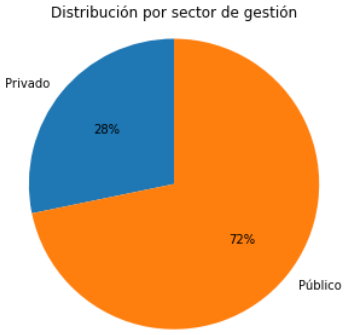


Figura [1]. Casi tres cuartas partes de la población relevada estudia en instituciones del sector público.

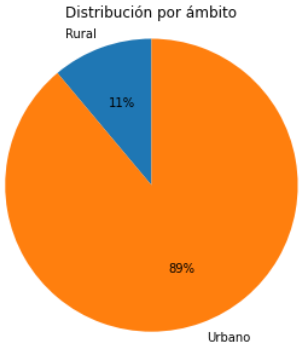


Figura [2]. La gran mayoría de los estudiantes corresponde a escuelas del ámbito urbano. Una décima parte corresponde al ámbito rural.

Figura [3]. Una décima parte de los estudiantes pertenece a hogares indígenas.

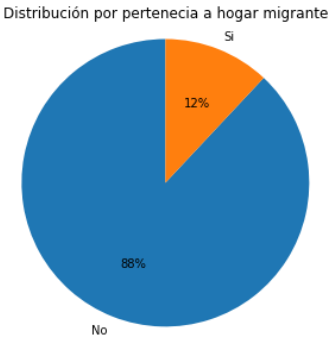


Figura [4]. Alrededor de una décima parte de los estudiantes pertenece a un hogar migrante.

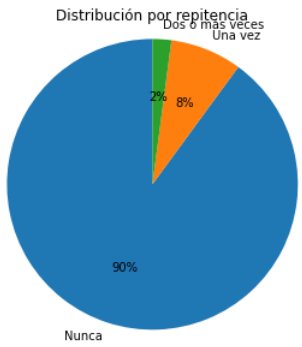


Figura [5]. La gran mayoría de los estudiantes no repitió de grado.

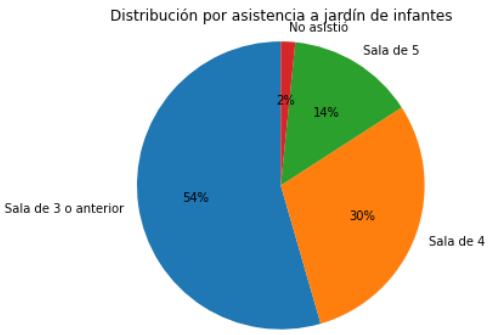


Figura [6]. La gran mayoría de los estudiantes asistió a jardín de infantes, siendo algo mayor la proporción de los que asistieron a sala de 3 o anterior vs los que asistieron solo a sala de 4 o 5.

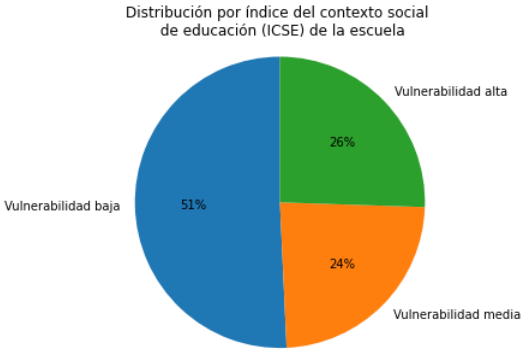


Figura [7][1]. La mitad de los estudiantes corresponde a escuelas de baja vulnerabilidad, mientras aproximadamente una cuarta parte corresponde a escuelas con vulnerabilidad media o alta.

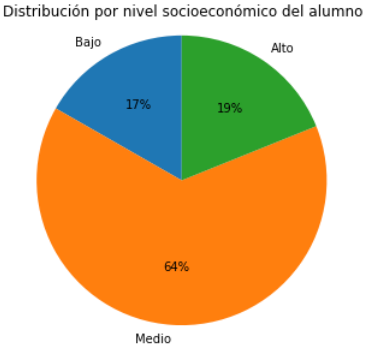


Figura [8]. Dos tercios de los estudiantes tiene un nivel socioeconómico medio, mientras que aproximadamente una sexta parte corresponde a un nivel bajo o alto.

Figura [9][2]. Casi la mitad de los estudiantes dice trabajar, ya sea para un familiar o alguien ajeno a la familia.

4.2. DISTRIBUCIÓN DE PUNTAJES GENERAL Y SEGÚN VARIABLES SELECCIONADAS

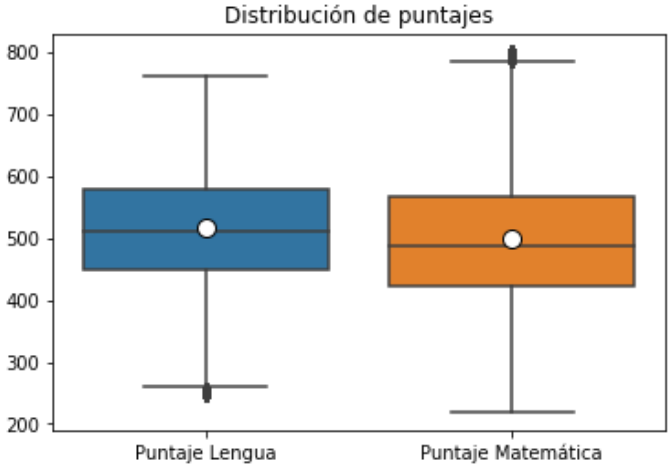


Figura [10]. Distribución de puntajes de lengua y matemática.

	Lengua	Matemática
Media	515	498
Mediana	511	486
Desvío estándar	86	101
Mínimo	250	216
Máximo	760	800

Tabla [1]. Medidas de tendencia central de los puntajes de lengua y matemática

[1] Esta pregunta contó con un porcentaje significativo de respuestas en blanco (75%)

[2] Variable propia que combina los resultados de las preguntas ap14 y ap15.

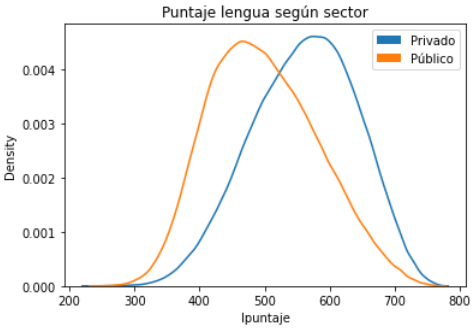


Figura [11]. Observamos que hay un desempeño marcadamente superior en el sector privado respecto al público.

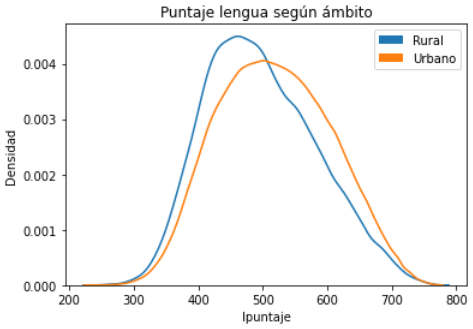


Figura [12]. Desempeño algo superior en el ámbito urbano vs rural.

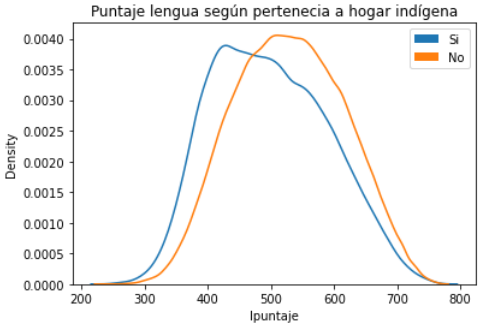


Figura [13]. Desempeño algo superior de estudiantes no pertenecientes a hogares indígenas vs sí pertenecientes.

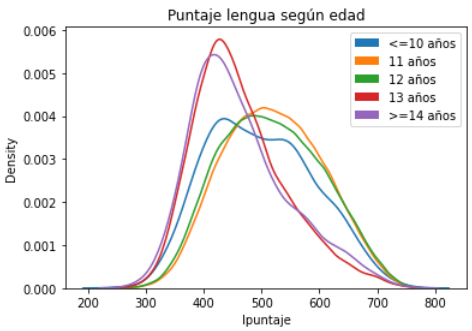


Figura [14]. Los estudiantes de 11 y 12 años se comportan de forma muy similar. Hay una proporción mínima de estudiantes de menos de 10 años con un desempeño algo inferior, lo que se acentúa en el caso de los estudiantes de 13 y 14 años (que podemos asociar a un fenómeno de repitencia).

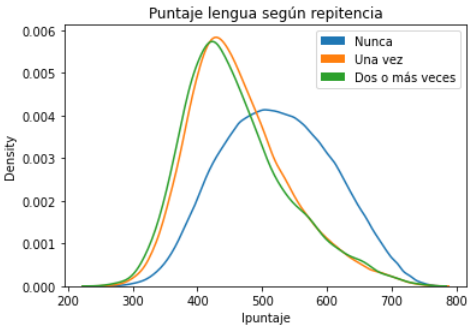


Figura [15]. Hay una diferencia bastante marcada entre aquellos que no repitieron y los que repitieron una o más veces.

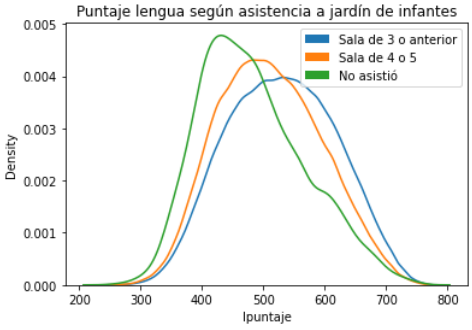


Figura [16]. Hay una diferencia bastante marcada entre los puntajes según si hubo o no asistencia a jardín de infantes desde una temprana edad

En la figura [10] y la tabla [1] se puede apreciar que los puntajes de lengua tienen una media y mediana algo mayor a los de matemática, a la vez que un desvío estándar menor. Los valores de las medias (515 para lengua y 498 para matemática) revelan una leve mejora y una mínima caída en los puntajes, respectivamente, en relación a la evaluación de referencia normalizada con media 500 (Aprender 2016). Además, el rango es un poco más acotado (510 puntos versus 584). Por otro lado, la correlación entre ambas variables es moderada, con un valor de 0,63.

En las figuras [11] a [16] las diferencias en los puntajes de lengua según algunas variables de interés: sector, ámbito, pertenencia a hogar indígena, edad, repitencia y asistencia a jardín de infantes. Estas visualizaciones nos sugieren cuales podrían ser las variables más útiles a la hora de entrenar un modelo predictivo que estime el puntaje de lengua.

5. PREPROCESAMIENTO

5.1. INGENIERÍA DE VARIABLES (*FEATURE ENGINEERING*)

Se crearon las siguientes variables, atendiendo a temáticas en común de diversos conjuntos de preguntas[3]:

Variable	Etiqueta	Dimensión latente	Variable creada
ap7a	¿Cuáles de estas cosas hay en el lugar donde vivís?... Conexión a Internet	Nivel de características del hogar	nivel_carac_hogar
ap7b	¿Cuáles de estas cosas hay en el lugar donde vivís?... Agua potable		
ap7c	¿Cuáles de estas cosas hay en el lugar donde vivís? Computadora		
ap7d	¿Cuáles de estas cosas hay en el lugar donde vivís? Heladera con freezer		
ap7e	¿Cuáles de estas cosas hay en el lugar donde vivís? Aire acondicionado		
ap7f	¿Cuáles de estas cosas hay en el lugar donde vivís? Calefacción		
ap13a	¿Con qué frecuencia te ocupás de las siguientes actividades en tu casa? Cuido a algún hermano u otro familiar	Nivel de actividad en el hogar	nivel_act_hogar
ap13b	¿Con qué frecuencia te ocupás de las siguientes actividades en tu casa? Realizo tareas del hogar como cocinar, limpiar, lavar la ropa, hacer las compras, etc.		
ap13c	¿Con qué frecuencia te ocupás de las siguientes actividades en tu casa? Cultivo, cosecho en la huerta, trabajo la tierra o cuido animales de granja para consumir en casa.		
ap25a	Por favor indicá cuántas veces los estudiantes de tu escuela...Molestan a los que se sacan buenas notas	Nivel de bullying	nivel_bullying
ap25b	Por favor indicá cuántas veces los estudiantes de tu escuela...Molestan a los que les va mal o repitieron		
ap25c	Por favor indicá cuántas veces los estudiantes de tu escuela...Discriminan por la religión		
ap25d	Por favor indicá cuántas veces los estudiantes de tu escuela...Discriminan por los aspectos físicos		
ap25e	Por favor indicá cuántas veces los estudiantes de tu escuela...Discriminan por tener alguna discapacidad		
ap25f	Por favor indicá cuántas veces los estudiantes de tu escuela...Discriminan por la nacionalidad		
ap25g	Por favor indicá cuántas veces los estudiantes de tu escuela...Dañan las cosas de la escuela		
ap25h	Por favor indicá cuántas veces los estudiantes de tu escuela...Insultan, amenazan o agreden a otros compañeros por redes sociales		
ap26a	Pensado en el último mes, ¿hiciste alguna de estas actividades en tu tiempo libre, fuera del horario escolar? Hice deporte	Nivel de actividades extraescolares	nivel_actividades
ap26b	Pensado en el último mes, ¿hiciste alguna de estas actividades en tu tiempo libre, fuera del horario escolar? Leí un libro		
ap26c	Pensado en el último mes, ¿hiciste alguna de estas actividades en tu tiempo libre, fuera del horario escolar? Me reuní con amigos		
ap26d	Pensado en el último mes, ¿hiciste alguna de estas actividades en tu tiempo libre, fuera del horario escolar? Estudié un idioma		
ap26e	Pensado en el último mes, ¿hiciste alguna de estas actividades en tu tiempo libre, fuera del horario escolar? Fui a ver algún espectáculo o exposición (cine, recital, teatro, museo...)		
ap28a	En tu escuela, ¿los docentes les hablaron de estos temas? Los cambios del cuerpo en la adolescencia	Nivel de educación sexual	nivel_eds
ap28b	En tu escuela, ¿los docentes les hablaron de estos temas? El cuidado del cuerpo y la salud		
ap28c	En tu escuela, ¿los docentes les hablaron de estos temas? El embarazo		
ap28d	En tu escuela, ¿los docentes les hablaron de estos temas? Métodos de prevención del embarazo y enfermedades de transmisión sexual		
ap28e	En tu escuela, ¿los docentes les hablaron de estos temas? Los derechos de niños, niñas y adolescentes		
ap28f	En tu escuela, ¿los docentes les hablaron de estos temas? Igualdad de derechos entre mujeres y varones		
ap28g	En tu escuela, ¿los docentes les hablaron de estos temas? Diversidad de las personas: apariencia física, orientación sexual e identidad de genero		
ap28h	En tu escuela, ¿los docentes les hablaron de estos temas? Prevención del maltrato		
ap28i	En tu escuela, ¿los docentes les hablaron de estos temas? Como evitar el abuso sexual		
ap28j	En tu escuela, ¿los docentes les hablaron de estos temas? Cuándo pedir ayuda a una persona de confianza		
ap28k	En tu escuela, ¿los docentes les hablaron de estos temas? La importancia de comunicar tus ideas		
ap28l	En tu escuela, ¿los docentes les hablaron de estos temas? La importancia del buen trato en la escuela		
ap20	¿Te va bien en tu clase de Lengua?	Nivel de autoevaluación en lengua	nivel_autoev_l
ap21a	En tu opinión, ¿cómo leés?		
ap21b	En tu opinión, ¿cómo escribís?		
ap22	¿Te va bien en tu clase de Matemática?	Nivel de autoevaluación en matemática	nivel_autoev_m
ap23	En tu opinión, ¿cómo resolvés los problemas de Matemática?		
ap14	Además de asistir a la escuela, ¿ayudás a tus padres o familiares en su trabajo?	Trabaja	trabaja
ap15	¿Trabajás fuera de tu casa para alguien que no sea parte de tu familia?		
ap9	¿Cuál es el máximo nivel educativo de tu mamá?	Máximo nivel educativo alcanzado por los padres	max_nivel_educativo_padres
ap10	¿Cuál es el máximo nivel educativo de tu papá?		

Tabla [2]. Variables creadas.

Las preguntas asociadas a las dimensiones nivel de actividad en el hogar, nivel de bullying, nivel de autoevaluación en lengua y nivel de autoevaluación en matemática contienen respuestas asociadas a una escala de Lickert, del tipo[4]:

1. Siempre
2. Muchas veces
3. Pocas veces
4. Nunca

[3] Para las cinco primeras dimensiones, estas pudieron ser identificadas por el número de pregunta, mientras que el detalle específico es identificado por la letra. En el caso de las últimas cuatro dimensiones, se utilizó un criterio propio.

[4] Ver anexo metodológico para más detalles

Es por esto que tomamos la decisión de utilizar la mediana del conjunto de respuestas para caracterizar la dimensión. Cabe destacar que debido a esto, las nuevas variables contarán con 7 pseudocategorías: a las 4 iniciales (1, 2, 3 y 4) se sumarán las intermedias (1,5; 2,5 y 3,5), fruto del cálculo de la mediana.

En el caso de las dimensiones nivel de activdades extraescolares y nivel de educación sexual, las variables involucradas son de tipo dicotómicas (0: No, 1: Sí). Es por esto que optamos por sumar las respuestas positivas. Por ejemplo, un valor de 3 para la dimensión "nivel de actividades extraescolares" significa que el estudiante ha realizado 3 actividades en el último mes.

Por otro lado, creamos la variables:

- Trabaja (trabaja), que simplemente combina los resultados de las variables asociadas al trabajo del estudiante, ya sea en el contexto familiar o no.
- Máximo nivel educativo de los padres (max_nivel_educativo_padres): calculada como Max(max_nivel_educativo_padre, max_nivel_educativo_madre)

Los valores nulos en las nuevas variables creadas fueron imputados con la mediana.

Por otra parte, previamente a ajustar los modelos se realizó una estandarización de los variables en puntaje z, de tipo:

$$x = \frac{(x - \bar{x})}{S}$$

Cabe aclarar que esta estandarización se realizó primero a partir del conjunto de entrenamiento, para luego rescalar las variables en el conjunto de prueba siguiendo el mismo promedio y desvío estándar (es decir, para evitar filtrar información del conjunto de prueba en el de entrenamiento).

5.2. RELACIÓN ENTRE NUEVAS VARIABLES Y PUNTAJE LENGUA

Pasamos a analizar las relaciones entre las nuevas variables creadas y el puntaje obtenido en lengua. En las figuras [17] a [19] se aprecian los promedios del puntaje según el valor del nivel, para los tres niveles con mayores niveles de correlación:

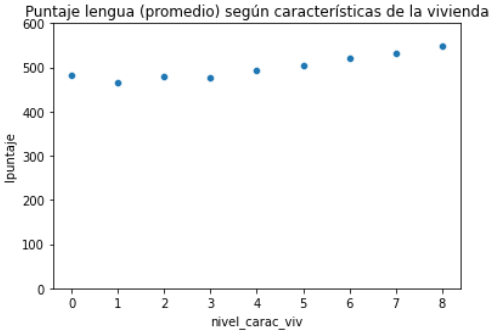


Figura [17]. Puntaje lengua según nivel de características de la vivienda

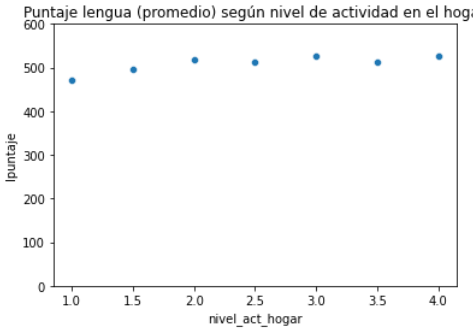


Figura [18]. Puntaje lengua según nivel de act. en el hogar

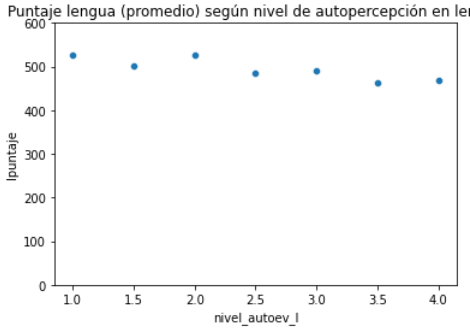


Figura [19]. Puntaje lengua según nivel de autopercepción en lengua

A continuación, calculamos el coeficiente de correlación de Spearman[5], para evaluar el grado de correlación entre cada variable y el puntaje:

[5] Utilizamos este test por considerar a los niveles como variables ordinales.

Variable	Coeficiente de correlación de Spearman	p-valor
Nivel de características de la vivienda	0,31	<0,01
Nivel de actividad en el hogar	0,11	<0,01
Nivel de bullying	0,07	<0,01
Nivel de actividades extraescolares	0,04	<0,01
Nivel de educación sexual	-0,04	<0,01
Nivel de autopercepción en matemática	-0,1	<0,01
Nivel de autopercepción en lengua	-0,12	<0,01

Tabla [3]. Niveles de correlación entre variables "Nivel" y puntaje lengua.

En todos los casos el coeficiente nos muestra grados de correlación bajos. Se destacan las variables **nivel de características de la vivienda**, con una correlación positiva, y el **nivel de autopercepción en lengua**, con una correlación ligeramente negativa. En el primer caso, podemos suponer que las caracterísiticas de la vivienda funcionan a modo de *proxy* del nivel socioeconómico del alumno. En el segundo caso, debe recordarse que mayores niveles numericos están asociados a una peor autopercepción, por lo que el resultado es el esperado.

6. REDUCCIÓN DE DIMENSIONALIDAD

6.1. JUSTIFICACIÓN TEÓRICA

A menudo los conjuntos de datos con múltiples variables, como el analizado en el este trabajo, contienen variables fuertemente correlacionadas, que aportan poca información nueva. Por otro lado, es inviable analizar este tipo de datasets de alta dimensionalidad a partir de diagramas de dispersión bivariados. Es por ello que queremos hallar una representación de baja dimensionalidad de los datos, que capture la mayor información posible. El análisis de componentes principales (PCA, por sus siglas en inglés) hace esto, al buscar un pequeño número de dimensiones que maximizen la varianza o inercia explicada. Así es como cada dimensión resultante -o componente principal- aparecerá como una combinación lineal de variables. Este método, sin embargo, es útil cuando se dispone solamente de variables continuas.

En el caso de contar con variables categóricas, podemos recurrir al análisis de correspondencias (CA), y su extensión para múltiples variables, el análisis de correspondiencias múltiples (MCA). Al tratarse variables categóricas, podemos calcular indicadores que hablen de la existencia o intensidad de la relación entre ellas, pero para conocer la naturaleza de esta relación debemos contemplar las categorías de dichas variables. Como indica Chan (2018), el procedimiento CA busca una representación en coordenadas de las filas y columnas de una tabla de contingencia, de modo tal que los patrones de asociación presentes en la tabla se reflejen en dichas coordenadas. Más específicamente, queremos resumir la información presente en las filas y columnas de manera que pueda proyectarse sobre un subespacio reducido, y representarse simultáneamente los puntos fila y los puntos columna, pudiéndose obtener conclusiones sobre relaciones entre las dos variables nominales u ordinales consideradas. Nuestro objetivo es proyectar estos puntos sobre un espacio de dimensión menor de manera tal que las filas que tengan estructuras similares aparezcan próximas y las que tengan estructuras diferentes aparezcan alejadas. Puede sintetizarse el procedimiento en los siguientes pasos:

1. Construimos las frecuencias relativas condicionales, consideradas como puntos del espacio.
2. Definimos la distancia Chi cuadrado entre estos puntos.
3. Proyectamos los puntos en el espacio que maximiza la variabilidad de la proyección.

Podemos extender esta lógica a múltiples variables a partir de la matriz disyuntiva o matriz de Burt, que resulta de la tabla de contingencia de todas las categorías o modalidades de las variables categóricas. La idea será resumir un conjunto de variables cualitativas utilizando variables cuantitativas asociando un coeficiente a cada categoría, y, para cada individuo, calculando la suma de los coeficientes de las categorías que posee.

Aclaración: a fin de poder aplicar el análisis MCA en nuestro conjunto de datos, tuvimos que realizar algunas transformaciones en las variables numéricas, a fin de categorizarlas. La estrategia utilizada fue dividir los datos en terciles

cuando fuera posible (es decir, en niveles bajo, medio y alto), o bien en dos niveles (bajo y alto), cuando un solo valor representara más de 33% de los casos (ver tabla [17] en anexo metodológico).

6.2. RESULTADOS

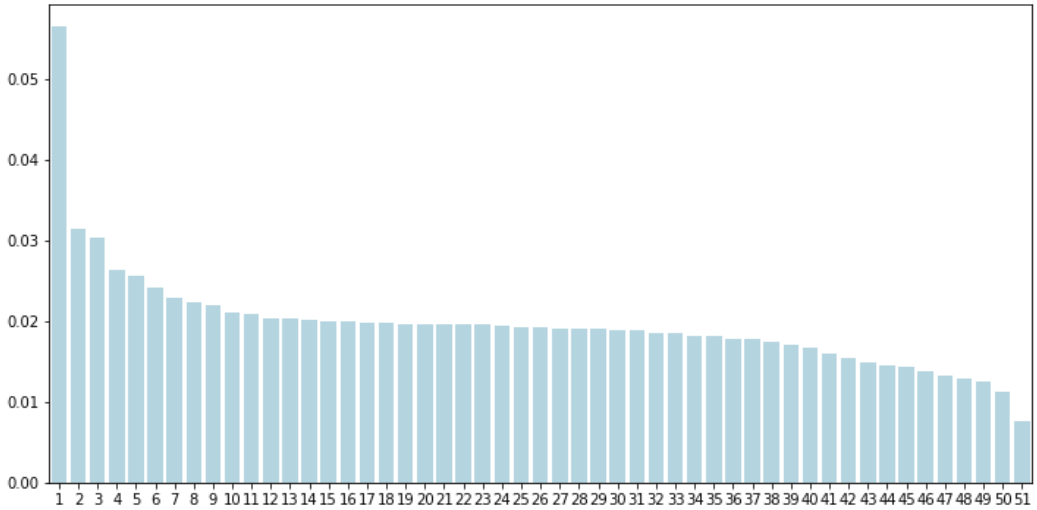


Figura [20]. Inercia explicada según los componentes principales (se omiten 22 componentes con valores muy cercanos a 0).

Variable	Nombre variable	Coeficiente
Edad: 14 años o más	edad_5	2,27
Repitió 2 veces	repitio_3	2,25
Repitió 3 o más veces	repitio_4	1,93
Edad: 13 años	edad_4	1,75
Máximo nivel educativo padres: sin educación	max_nivel_ed_padres_1	1,61
¿Fuiste a jardín de infantes? No	asistencia_jardin_4	1,58
Máximo nivel educativo padres: primaria incompleta	max_nivel_ed_padres_2	1,38
Repitió 1 vez	repitio_2	1,32
Ámbito: rural	ambito_0	1,07
Máximo nivel educativo padres: primaria completa	max_nivel_ed_padres_3	0,99
Sector: privado	sector_0	-0,93
Aproximadamente, ¿cuántos libros hay donde vivís? No hay	cant_libros_viv_1	0,91
¿Cuántas personas viven en tu casa? 9	cant_personas_viv_9	0,89
Nivel de cataterísticas de vivienda: alto	nivel_carac_viv_2	-0,84
¿Cuántas personas viven en tu casa? 8	cant_personas_viv_8	0,81

Tabla [4]. Coeficientes de primer componente principal.

Variable	Nombre variable	Coeficiente
Repitió 2 veces	repitio_3	0,95
Edad: 14 años o más	edad_5	0,93
¿Te gusta ir a la escuela? No	gusta_escuela_0	-0,87
¿Te llevás bien con tus compañeros y compañeras? Nunca	buena_relacion_comp_4	-0,82
Nivel autoevaluación matemática: medio	nivel_autoev_l_1	-0,81
Nivel autoevaluación matemática: bajo	nivel_autoev_m_2	-0,77
Ámbito: rural	ambito_0	0,75
Hogar indígena: sí	hogar_indigena_1	0,67
Nivel autoevaluación matemática: alto	nivel_autoev_m_0	0,63
¿Fuiste a jardín de infantes? No	asistencia_jardin_4	0,58
¿Te llevás bien con tus compañeros y compañeras? Pocas veces	buena_relacion_comp_3	-0,58
Edad: 13 años	edad_4	0,54
Repitió 3 o más veces	repitio_4	0,54
¿Te llevás bien con tus compañeros y compañeras? Siempre	buena_relacion_comp_1	0,53
Máximo nivel educativo padres: primaria incompleta	max_nivel_ed_padres_2	0,53

Tabla [5]. Coeficientes de segundo componente principal.

Según la figura [20], podemos apreciar que los primeros 10 componentes explican el 28% de la inercia o variabilidad total de los datos. En particular, nos concentraremos en los primeros 2 componentes (que sumados explican el 9% de la inercia), ya que serán los que podremos graficar, para así obtener conclusiones sobre la interacción entre las modalidades de las diferentes variables.

Podemos interpretar a los ejes como factores ocultos o variables latentes. Como vemos en la tabla [4], el primer componente principal indica en que medida los estudiantes son repetidores, de edad mayor a la esperada, con padres con poca educación, del ámbito rural y sector público, que viven en condiciones de hacinamiento y no asistieron al jardín de infantes. Por otro lado, el segundo componente (tabla [5]), tiene elementos en común al primero (repetidores de edad avanzada), pero se asocia además a la medida en que se trata de hogares indígenas y rurales, con niños que gustan de ir a la escuela, se llevan bien con los compañeros y tienen un nivel de autopercepción en matemática elevado.

Podemos obtener algunas apreciaciones a partir de las coordenadas de los componentes, tal como se observa en la figura [21]:

- Un máximo nivel educativo de los padres bajo (sin educación formal o con escuela primaria incompleta) se asocia a niños repetidores, que no fueron al jardín, y con edad mayor a la correspondiente a 6º grado.Además, que viven en condiciones de hacinamiento.
- Niveles bajos de autopercepción en matemática y lengua se asocian a niños que se llevan mal con los compañeros.
- Hogares indígenas se asocian al ámbito rural.

Por otro lado, podemos apreciar que ambos componentes parece ser útiles al momento de diferenciar los niveles de desempeño en lengua, en tanto hay una clara correlación negativa (los niveles de desempeño se reducen en la medida en que pasamos del cuadrante inferior izquierdo al cuadrante superior derecho).

A modo de ejemplo, podemos caracterizar algunos casos extremos del primer componente:

- El alumno 197892, con un valor mínimo en el componente 0, estudia en un entorno urbano y en el sector privado, viven 3 personas en su hogar, sus padres tienen estudios universitarios completos, fue al jardín antes de los 4 años y no repitió de grado. Tiene el máximo nivel de desempeño en lengua.
- Por otro lado, el alumno 293615, con un valor máximo en el mismo componente, pertence a un hogar indígena, estudia en un entorno rural en el sector público, en condiciones de hacinamiento (8 personas en pocas habitaciones), con padres sin educación formal. No fue al jardín y repitió 2 veces. Tiene un nivel desempeño mínimo en lengua.

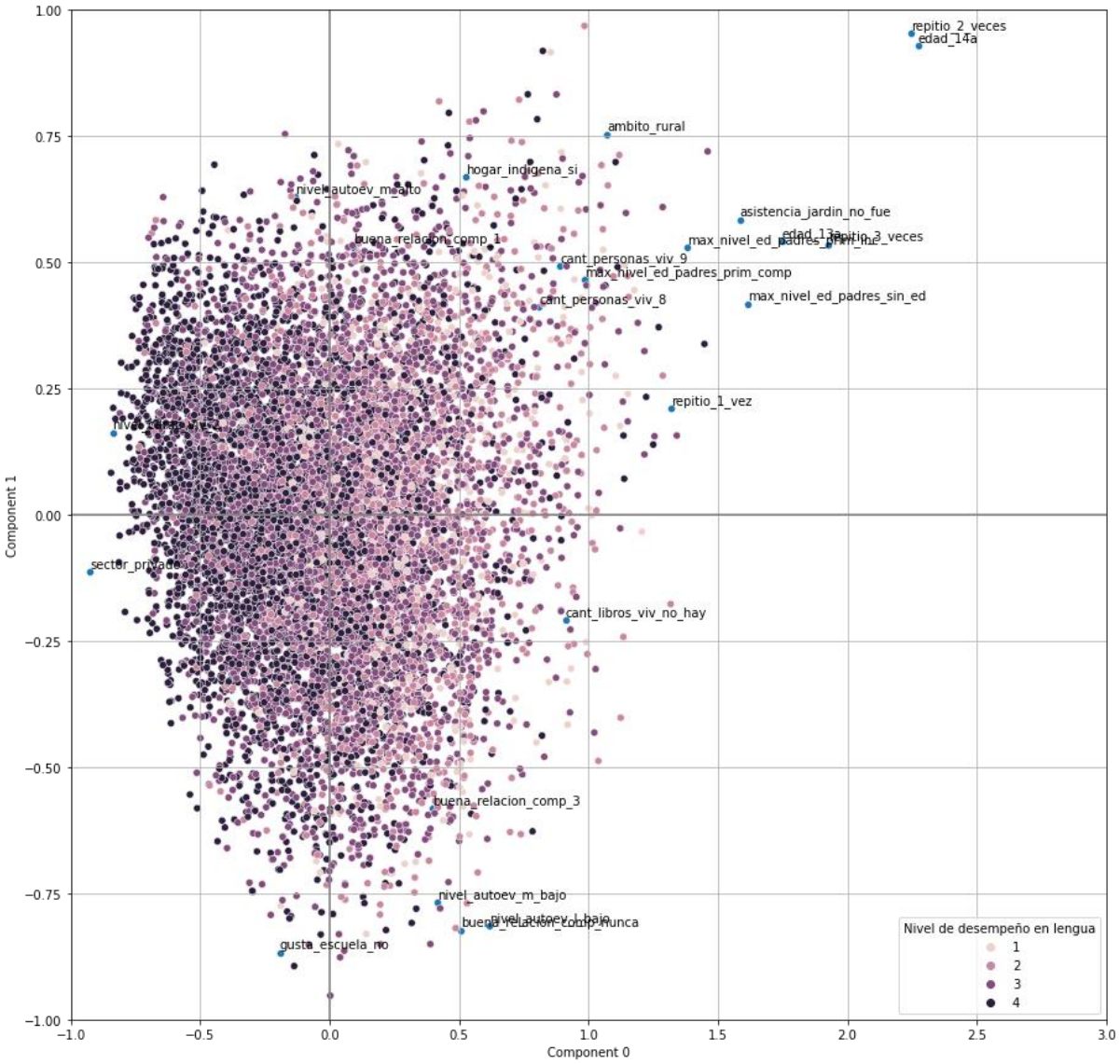


Figura [21]. Biplot con filas y columnas de MCA. Para facilitar la visualización e interpretación, se tuvieron en cuenta las 22 categorías con mayores coeficientes absolutos de los primeros 2 componentes.

7. MODELOS

7.1. ESQUEMA DE VALIDACIÓN

Dividimos los datos en un conjunto de entrenamiento y otro de prueba, según el siguiente esquema:

Conjunto	Porcentaje de registros	Cantidad de registros
Entrenamiento	80%	432.550
Prueba	20%	108.138
TOTAL	100%	540.688

Tabla [6]. División en conjuntos de entrenamiento y prueba.

A fin de reducir los tiempos de procesamiento, utilizamos un muestreo representativo con 100.000 casos, aproximadamente un 20% del total de registros, a fin de realizar pruebas con diferentes configuraciones de modelos. Es decir, tenemos el siguiente esquema:

Conjunto	Porcentaje de registros	Cantidad de registros
Entrenamiento	80%	80.000
Prueba	20%	20.000
TOTAL	100%	100.000

Tabla [7]. División en conjuntos de entrenamiento y prueba. Submuestro para pruebas.

El conjunto de variables utilizadas para entrenar (en su estado original, es decir, sin transformaciones), sumadas a las variables agregadas en el proceso de *feature engineering* es de 104. Una vez consumadas las transformaciones ya reseñadas, el número de variables asciende a 235.

Por otro lado, armamos un esquema de validación cruzada (*cross-validation*, en adelante CV), al momento de buscar los mejores hiperparámetros de los diferentes algoritmos entrenados, a fin de evitar el sobreajuste a los datos de entrenamiento, para así contar con un modelo más robusto, que tenga validez para ser aplicado en datos desconocidos. Este esquema se desarrolla de la siguiente manera:

1. Dividimos el conjunto de entrenamiento en 5 partes o *folds* iguales (1, 2, 3, 4, 5).
2. Entrenamos un modelo en los primeros 4 folds (1, 2, 3, 4).
3. Testeamos el modelo entrenado en el fold restante (5).
4. Entrenamos un nuevo modelo, ahora con los folds 2, 3, 4 y 5, y lo testeamos con el fold restante (1).
5. Iteramos 5 veces el proceso anterior, de forma tal de entrenar 5 modelos distintos con la combinación de 4 folds, puestos a prueba con cada uno de los folds restantes.
6. Obtenemos el promedio de la métrica obtenida en todas las iteraciones.

Utilizamos la métrica **RMSE** (Root Mean Square Error o Error Cuadrático Medio), que mide la raíz del promedio de los errores entre el valor estimado y el valor real. En nuestro caso, tendremos en cuenta el puntaje de lengua estimado vs el real.

7.2. AJUSTE DE MODELOS[6]

7.2.1. MODELO: LÍNEA DE BASE (*BASELINE*)

Al momento de ajustar modelos predictivos sobre nuestro conjunto de datos, siempre es conveniente tener un modelo de referencia contra el cual comparar los resultados, y así poder establecer niveles de mejora contra un nivel de línea de base (o *baseline*). En nuestro caso, podemos pensar en tomar simplemente el promedio del puntaje del

[6] Cabe aclarar que, tal como fuera mencionado en la introducción, en el entrenamiento de los modelos se utilizó el ponderador *lpondera,* a fin de darle un peso diferenciado a cada registro.

conjunto de entrenamiento como dicho *baseline*: una predicción del puntaje sin tomar en consideración ninguna variable predictora. Esta es la misma metodología que utilizan Cortes y Silva [2].

7.2.2. MODELO: REGRESIÓN LINEAL

7.2.2.a. DEFINICIÓN

Podemos definir el modelo de regresión lineal simple tal como lo hace Szretter [5]:

> El modelo de regresión lineal es un modelo para el vínculo de dos variables aleatorias que denominaremos X = variable predictora o covariable e Y = variable dependiente o de respuesta. El modelo lineal (simple pues sólo vincula una variable predictora con Y) propone que:
>
> $$y_i = \beta_0 + \beta_1 x + \varepsilon$$
>
> donde ε es el término del error. Esto es que para cada valor de X, la correspondiente observación Y consiste en el valor $\beta_0 + \beta_1 x$ más una cantidad ε que puede ser positiva o negativa, y que da cuenta de que la relación entre X e Y no es exactamente lineal, sino que está expuesta a variaciones individuales que hacen que el par observado (X; Y) no caiga exactamente sobre la recta, sino cerca de ella.
> Los números $\beta_0 + \beta_1$ son constantes desconocidas que se denominan parámetros del modelo, o coecientes de la ecuación. El modelo se denomina lineal pues propone que la Y depende linealmente de X. Además, el modelo es lineal en los parámetros: los β s no aparecen como exponentes ni multiplicados o divididos por otros parámetros. Los parámetros se denominan:
>
> $$\beta_0 = ordenada\ al\ origen$$
> $$\beta_1 = pendiente$$

La meta del modelo de regresión lineal será obtener los parámetros β_0 y β_1 que mejor se ajusten a los datos disponibles. El método más común para realizar esto es el de mínimos cuadrados.

Siendo: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la predicción para Y en base al iésimo valor de X, entonces $e_i = y_i - \hat{y}_i$ representa al iésimo error: la diferencia entre el iésimo valor de respuesta observado y el iésimo valor predicho por nuestro modelo. Definimos entonces la suma de cuadrados como: RSS = $e_0^2 + e_1^2 + \cdots + e_n^2$.
Expresado de otra manera:

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

Con el método de mínimos cuadrados buscamos encontrar β_0 y β_1 de forma tal de minimizar el valor de RSS:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

El modelo de regresión lineal múltiple aparece como una extensión natural de este modelo, al incorporar más de una variable predictora: $X_1, X_2, \ldots, X_{p-1}$, con sus respectivos coeficientes (o parámetros del modelo) $\beta_1, \beta_2, \ldots, \beta_{p-1}$.

7.2.2.b. EL PROBLEMA DE LA COLINEALIDAD

La colinealidad se refiere a la situación en que dos o más variables predictoras se encuentran estrechamente correlacionadas. Esto puede suponer problemas en un ejercicio de regresión lineal, por cuanto dificulta singularizar el aporte individual de cada variable en el resultado final, al incrementar el error estándar en la estimación de cada coeficiente beta. Una posible solución a este problema consiste en establecer una matriz de correlaciones, para entender las correlaciones de a pares.

Sin embargo, pueden ocurrir problemas de colinealidad entre 3 o más variables, sin que las variables tomadas de a pares muestren niveles altos de correlación. Este problema se denomina multicolinealidad. Una forma de solucionarlo es a partir del método VIF (Variance Inflation Factor o Factor de Inflación de la Varianza). El valor VIF consiste en el coeficiente de varianza de $\hat{\beta}_j$ al ajustar el modelo completo versus la varianza de $\hat{\beta}_j$ únicamente. En otros términos, podemos calcular VIF como:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

siendo el $R^2_{X_j|X_{-j}}$ resultado de calcular la regresión de X_j a partir del resto de las variables predictoras. En la medida en que R^2 se acerque a 1, el valor VIF se incrementará [6].

Como vimos, nuestro modelo cuenta mayoritariamente con variables categóricas que, al ser codificadas en variables *dummy*, significaron un incremento sustancial de la dimensionalidad (235 variables). Esta circunstancia supone una probabilidad mayor de tener problemas de multicolinealidad, que deben ser contemplados.

7.2.2.c. TÉCNICAS DE REGULARIZACIÓN: RIDGE Y LASSO

Podemos utilizar dos técnicas para reducir o regularizar los coeficientes: las técnicas de Ridge y Lasso. Por un lado, la técnica de Ridge busca minimizar la siguiente ecuación:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2,$$

siendo λ el parámetro de ajuste. El segundo término de la ecuación, o penalidad, tiene el efecto de reducir los coeficientes beta hacia 0. El parámetro λ controla el grado de penalidad: con un valor igual a 0 no tendrá ningún efecto y el resultado final sería idéntico a una regresión lineal estándar, mientras que en la medida en que se acerque a infinito el grado de penalidad se incrementará, y los coeficientes de regresión se acercarán a 0.

El parámetro λ incidirá sobre el equilibrio sesgo-varianza. Con λ = 0, la varianza es alta, pero no hay sesgo. En la medida en que λ se incremente, se reducirá sustancialmente la varianza al costo de un leve incremento en el sesgo. Como indica Tibshiriani [6], "en general, en situaciones en que la relación entre las variables predictoras y la respuesta es lineal, los [coeficientes] estimados por mínimos cuadrados tendrán poco sesgo y mucha varianza. (…) Esto significa que un pequeño cambio en los datos de entrenamiento puede significar en un gran cambio en los coefientes estimados por mínimos cuadrados."

La técnica de Lasso introduce una modificación respecto a Ridge. En esta última, con un λ suficientemente alto, los coeficientes tenderán a 0, pero nunca serán 0. Esto puede suponer un problema para la interpretación de los coeficientes de un número alto de variables. Lasso modifica la ecuación de Ridge, buscando minimizar la siguiente cantidad:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

La única diferencia reside en que, mientras que Ridge incluye la sumatoria de β_j^2, Lasso incluye la sumatoria de $|\beta_j|$. Este cambio supone la posibilidad de llevar algunos coeficientes a ser iguales a 0, ejecutando de esta manera un proceso de selección de variables, y ayudando así a la interpretabilidad del modelo. Por otro lado, esta técnica puede ayudar a reducir la multicolinealidad presente en el conjunto de datos, penalizando fuertemente los coeficientes de variables que están correlacionadas.

7.2.3. ÁRBOLES DE DECISIÓN

Los árboles de decisión implican estratificar o segmentar el espacio de predictores en cierto número de regiones más simples. Para hacer una predicción para determinada observación, típicamente utilizamos la media o la moda de las observaciones de entrenamiento de la región a la que pertenece. Dado que el conjunto de reglas de división utilizadas para el segmentar el espacio de predictores puede ser sintetizado en un árbol, este tipo de acercamiento es

conocido como métodos de árboles de decisión [6]. A su vez, una vez construido un árbol, uno puede, a posterior, extraer un conjunto de reglas del tipo SI, ENTONCES, que son de facil comprensión.

7.2.4. EXTREME GRADIENT BOOSTING (XGBOOST)

Como explica Tibshiriani [6], los árboles de decisión son procedimientos estadísticos con alta varianza: si dividimos un conjunto de entrenamiento en dos mitades al azar, y aplicamos un árbol de decisión a ambas partes, los resultados que obtendríamos serían bastante diferentes; por el contrario, un procedimiento con poca varianza generaría resultados similares al ser aplicado a diferentes conjuntos de datos (es el caso, por ejemplo, de la regresión lineal, con una proporción de n a p suficientemente grande). La agregación por *bootstrap* (*bagging*) permite reducir la varianza de un método de aprendizaje estadístico. Específicamente, esto significa tomar muchas muestras con repetición de la población (con el método *bootstrap*), construir un modelo diferente para cada uno y promediar los resultados. En el caso de los árboles de decisión, se entrenan modelos profundos sin proceso de *pruning* (o poda), lo que implica que tengan poco sesgo y mucha varianza, la cual será contrarrestada por el método de *bagging*, que ha demostrado producir notables incrementos en la precisión de las predicciones.

A diferencia del método de bagging, el método de boosting entrena los árboles de forma secuencial. A partir de un modelo inicial ajustamos subsecuentes modelos no a la variable objetivo *y*, sino a los residuos del modelo anterior. Estos modelos serán más bien pequeños y débiles (*weak learners*), e irán mejorando lentamente el ajuste en aquellas áreas en donde haya mayor error, actualizando de esta forma los parámetros del árbol. Se puede limitar el crecimiento de los árboles a partir de los hiperparámetros: cantidad de árboles (controla el sobreajuste del ensamble), parámetro de encogimiento o shrinkage λ (controla la tasa de aprendizaje), y la cantidad d de divisiones en cada árbol (controla la complejidad del ensamble).

Gradient boosting (o boosting por gradiente) aparece como una variante del método *boosting*, adonde el objetivo es minimizar la función de pérdida del modelo a partir del método de descenso del gradiente. La función de pérdida debe ser diferenciable, y podría utilizarse, en el caso de un problema de regresión, la ya reseñada métrica RMSE o bien MAE (el error absoluto medio).

Extreme Gradient Boosting (XGBoost) es un modelo de boosting desarrollado a partir de 2014 por Tianqui Chen que busca al mismo tiempo reducir los tiempos de ejecución y optimizar el desempeño del modelo de gradient boosting. Tiene como características salientes [1] [3]:

- El manejo inteligente de valores nulos (*sparsity aware*).
- La paralelización en la construcción de árboles (*column block*).
- El entrenamiento continuo, que permite continuar mejorando un modelo ya ajustado a un conjunto de datos.

XGBoost introduce parámetros como column subsampling (o submuestro de columnas), que, al igual que los bosques aleatorios o random forests, selecciona solo una muestra de las variables para ajustar cada nuevo árbol, lo que permite evitar el sobreajuste, al mismo tiempo que reduce los tiempos de procesamiento.

8. RESULTADOS

Algoritmo	RMSE promedio en CV	Tiempo de ejecución	Hiperparámetros
Baseline	87,2	00:00:02	-
Regresión lineal	68,7	00:00:23	-
Lasso	68,8	00:01:14	Lambda=0,1
Ridge	68,6	00:00:15	Lambda=100
Árbol de decisión	75,2	00:00:02	max_depth=8
XGBoost	69,3	00:03:37	Por defecto
XGBoost*	69,1	00:01:31	Por defecto
XGBoost*	67	00:01:51	Optimización bayesiana
XGBoost*	67,3	00:01:51	Optimización bayesiana con selección de variables (n=65)

Tabla [8] Desempeño de modelos.

**En estos modelos utilizamos la base original, sin desglose de variables categóricas en variables dummy (es decir, tomando la base original de 104 variables). Si bien no hay una explicación formal para este fenómeno, en nuestra experiencia el algoritmo XGBoost logra un desempeño algo superior (además de suponer un menor tiempo de entrenamiento), quizás interpretando las variables categóricas como multimodales.*

8.1. MODELOS LINEALES

Podemos apreciar que los modelos de regresión lineal -en sus diferentes variantes- suponen una mejora significativa del 21% en la métrica RMSE, al compararlos contra el modelo baseline (promedio del puntaje en el conjunto de entrenamiento).

Al momento de analizar los modelos lineales, encontramos que el modelo de regresión lineal ajustado arroja coeficientes muy altos, no interpretables , por lo que calculamos el índice VIF. Este nos indica el 60% de las variables del conjunto del entrenamiento tienen un valor VIF >= 5, por lo que asumimos un alto nivel de multicolinearidad.

Consideramos apropiada la utilización del modelo Lasso, por cuanto:

1. Penaliza los coeficientes altos, permitiendo una interpretación apropiada.
2. Aplica un proceso de selección de variables, simplificando el modelo y evitando el sobreajuste. Puntualmente, el modelo descarta 87 variables predictoras (llevando sus coeficientes a 0), dejando un total de 148 variables con coeficientes con valores absolutos positivos.
3. El desempeño del modelo Lasso con el parámetro lambda optimizado es prácticamente indistinguible de aquel del modelo de regresión básico o el Ridge.

En la tabla [9], podemos ver los 5 coeficientes más altos y los 5 más bajos de la regresión Lasso[7]. La ordenada al origen es igual a 497. Vale decir, con todas las variables predictoras en 0 tendremos un puntaje de lengua base de 497 puntos.

Por ejemplo, podemos interpretar los coeficientes de la siguiente forma:

Variable	Pregunta de referencia	Respuesta categórica	Coef.
bullying_religion_4	Por favor indicá cuántas veces los estudiantes de tu escuela…Discriminan por la religión	4. Nunca	13
act_hogar_tareas_2	¿Con qué frecuencia te ocupás de las siguientes actividades en tu casa? Realizo tareas del hogar como cocinar, limpiar…	2. Muchas veces	13
bullying_fisico_2	Por favor indicá cuántas veces los estudiantes de tu escuela…Discriminan por los aspectos físicos	2. Muchas veces	12
autoevaluacion_lectura_1	En tu opinión, ¿cómo leés?	1. Muy bien	12
asistencia_jardin_1	¿Fuiste a jardín de infantes?	1. Sí, fui al jardín antes de los cuatro años	10
cant_libros_viv_1	Aproximadamente, ¿cuántos libros hay donde vivís?	De 1 a 25 libros	-11
autoevaluacion_lengua_4	¿Te va bien en tu clase de Lengua?	4. No muy bien	-11
trabaja_sin_familia	¿Trabajás fuera de tu casa para alguien que no sea parte de tu familia?	0. No, 1. Sí	-13
autoevaluacion_escritura_1	En tu opinión, ¿cómo escribís?	1. Muy bien	-15
sector	Sector de gestión	0. Privado, 1. Público	-23

Tabla [9]. 5 coeficientes más altos y 5 coeficientes más bajos de regresión lineal con regularización Lasso.

La pertenencia al sector de gestión público, manteniendo las otras variables constantes, supone una reducción en el puntaje de lengua de 23 puntos, respecto de la línea de base de la ordenada al origen.

Hay algunas variables que pueden interpretarse con mayor facilidad, como ser la pertenencia al sector público (víctima de desfinanciación en las últimas décadas) o el hecho de que el niño trabaje, que implican una reducción del puntaje de lengua esperado, o bien la autoevaluación positiva en lectura o la asistencia a jardín de infantes desde temprana edad, que implican un incremento en el puntaje. El nivel de actividad en el hogar, con un coeficiente positivo, puede vincularse con la ya reseñada correlación de Spearman destacada. Llama la atención que la autoevaluación en escritura positiva suponga una reducción en el puntaje, pero esto puede estar ligado a una fuerte correlación de esta variable con el resto de las variables predictoras (puntaje VIF = 33), lo que impide aislar su efecto puntual sobre la variable a predecir. Por otro lado, el hecho de que pertenecer al ámbito urbano reduzca el puntaje no es un resultado esperado (como vimos, los puntajes para el ámbito urbano son mayores en promedio al ámbito rural), pero esto puede explicarse por algún fenómeno de interacción entre variables no contemplado en este modelo (que pueden ser objeto de futuros análisis).

[7] Modelo ajustado sin validación cruzada, sobre el total del subconjunto de 80.000 casos.

8.2. MODELOS DE ÁRBOLES

8.2.1. DESEMPEÑO

Como apreciamos en la tabla [8], el modelo 1 con árboles de decisión (con el hiperparámetro profundidad optimizado) representa una mejoría sobre el modelo de línea de base, pero no sobre los modelos de regresión. Por otro lado, el modelo XGBoost -con hiperparámetros por defecto- logra un desempeño claramente mejor al de un árbol solo, pero sigue siendo inferior al de los modelos de regresión. Sin embargo, al realizar una optimización bayesiana de los hiperparámetros, logramos una mejora de 1,8 puntos sobre el modelo Lasso. Esto supone además una mejora de 23% del RMSE sobre el modelo de línea de base.

Pasamos a detallar algunas apreciaciones adicionales sobre los modelos de árboles ajustados.

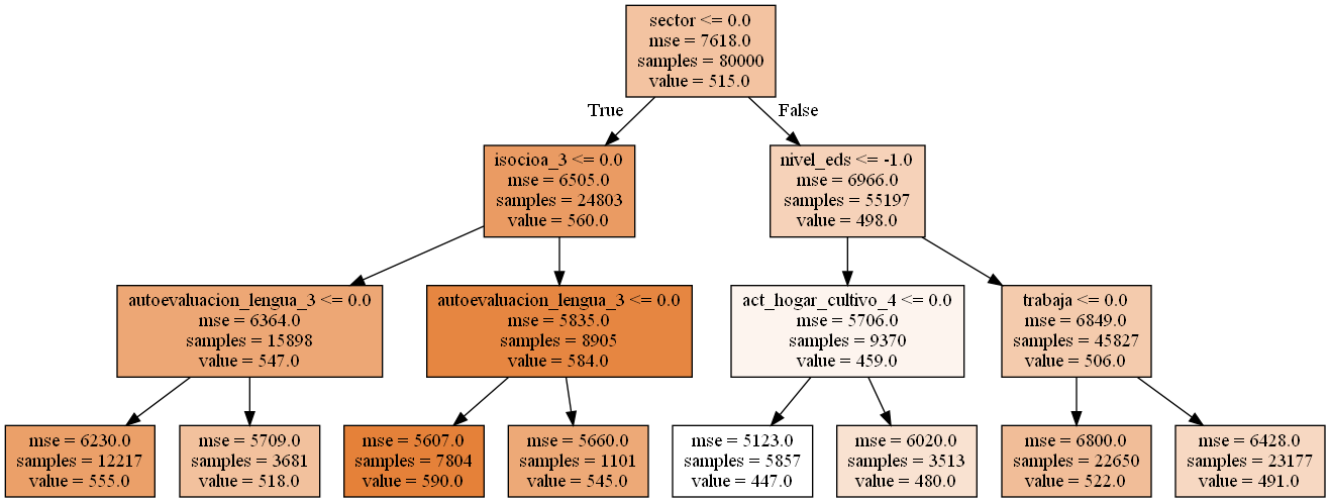


Figura [22]. Ejemplo de árbol de decisión (profundidad 3).

Como vemos en la figura [22], al igual que en el modelo de regresión, el sector aparece como la variable preponderante (por estar en la raíz del árbol). Hay algunas variables con una interpretación más clara, como ser el índice socioeconómico del alumno, la autoevaluación en lengua positiva (que puede vincularse con la ya reseñada correlación de Spearman destacada), el trabajo en actividades de cultivo en el seno del hogar o el hecho de trabajar o no. Por ejemplo, el mayor puntaje (590) correspondería a un estudiante del sector privado, con un índice socioeconómico alto y una autoevaluación en lengua distinta a regular (categoría 3). A su vez, el menor puntaje (447) correspondería a un estudiante del sector público, con un índice de educación sexual en la escuela bajo[8] y con cierto nivel de trabajo en actividades de cultivo en el hogar (categoría <=3).

[8] Tener en cuenta que esta variable está estandarizada en puntaje Z.

8.2.2. IMPORTANCIA DE VARIABLES XGBOOST (*FEATURE IMPORTANCE*)

Variable	Pregunta de referencia	Feature importance
sector	Sector	0,19
trabaja	Trabaja (ya sea en el marco de la familia o no)	0,05
isocioa	Indice socioeconómico del alumno	0,04
bullying_religion	Por favor indicá cuántas veces los estudiantes de tu escuela…Discriminan por la religión	0,04
carac_viv_pc	¿Cuáles de estas cosas hay en el lugar donde vivís? Computadora	0,03
eds_igualdad_genero	En tu escuela, ¿los docentes les hablaron de estos temas? Igualdad de derechos entre mujeres y varones	0,03
autoevaluacion_lengua	¿Te va bien en tu clase de Lengua?	0,03
nivel_carac_viv	Nivel características de la vivienda	0,03
act_hogar_tareas	¿Con qué frecuencia te ocupás de las siguientes actividades en tu casa? Realizo tareas del hogar como cocinar…	0,02
trabaja_con_familia	Además de asistir a la escuela, ¿ayudás a tus padres o familiares en su trabajo?	0,02

Tabla [10]. Importancia relativa de variables modelo XGBoost (optimizado). 10 variables más importantes. Valores normalizados (sumatoria = 1).

Como apreciamos en la tabla [10], la mayoría de las 10 variables más importantes para el modelo XGBoost coinciden con aquellas destacadas en el árbol de decisión único, o bien en la regresión lineal Lasso. De hecho, las únicas variables nuevas que encontramos son nivel_carac_viv, carac_viv_pc y eds_igualdad_genero. Una de las desventajas de los modelos de ensamble como el XGBoost, es que no podemos saber si las variables en cuestión afectan de forma positiva o negativa al modelo, a diferencia de los modelos de regresión. Es decir, ganamos *mayor poder predictivo a expensas de sacrificar intepretabilidad en el modelo*. Cabe destacar, por último, que la suma de la importancia de estas 10 variables (de un total de 104) totaliza 47%, lo que sugiere que muchas variables del conjunto de datos aportan muy poca capacidad explicativa al modelo.

Una alternativa para mesurar el aporte de las variables al modelo es a través de los valores SHAP (SHapley Additive exPlanations). Estos valores están inspirados en los valores Shapley de la teoría de juegos. La idea es realizar variaciones del modelo original, utilizando solo una parte de las variables predictoras, a fin de estimar la contribución marginal de cada variable al resultado final. Los valores SHAP poseen tres importantes características:

- Interpretabilidad global: muestran cuanto aporta cada predictor, ya sea positiva o negativamente, a la variable *target*.
- Intepretabilidad local: cada observación tiene su conjunto específico de valores SHAP. Esto puede ayudar a mostrar relaciones no lineales entre las variables predictoras y la variable *target*.
- Los valores SHAP pueden calcularse para cualquier método basado en árboles.

Como vemos en la figura [23], hay algunas diferencias con respecto a la tabla de importancia de variables original, calculada por defecto por el paquete XGBoost. Mientras que las variables sector, trabaja, bullying_religion, autoevaluacion_lengua y act_hogar_tareas siguen siendo importantes, aparecen nuevas variables entre las más destacadas, como ser cant_personas_viv, bullying_dañan_escuela, bullying_fisico y provincia.

En la figura [24] podemos observar la influencia de las variables más importantes, *desagregadas por cada caso particular.* Recordar que en el caso de las variables de autoevaluación y bullying *el sentido está invertido*, en tanto valores más bajos (1) corresponden a una muy buena autoevaluación y a la ocurrencia continua de bullying, y valores bajos (4) corresponden a una mala autoevaluación y la no ocurrencia de bullying.

- En algunas variables es unívoca la relación: el sector privado (0) está asociado a mayores puntajes, al igual que el hecho de no trabajar. La cantidad de personas en el hogar está correlacionada de forma negativa, mientras que mayores niveles de participación en tareas del hogar correlacionan de forma positiva.
- Por otro lado, mejores niveles conceptuales de autoevaluacion en lengua y lectura se asocian a mayores puntajes, lo que esperable. Este comportamiento ya había aparecido en el análisis de correlación de autoevaluación en lengua vs puntaje, y en el modelo Lasso. Debido a que los puntajes están invertidos, la correlación *negativa* de autoevaluacion_lengua y autoevaluacion_lectura deben interpretarse como una correlación *positiva*.

- Por último, se aprecia aquí la disímil influencia de las variables según el caso. Incluso, vemos relaciones no lineales. Por un lado, la mayor ocurrencia (1 o 2) de bullying_dañan_escuela y bullying_físico influyen tanto positiva como negativamente, pero en poca magnitud. En el caso de bullying_religion la mayor ocurrencia están asociada a un peor rendimiento. Por otro lado, la existencia baja o nula de bullying (niveles 3 o 4) parece influir de forma dísimil según el caso.

Es interesante observar lo que ocurre en casos individuales, en los que la predicción será igual a la suma de los valores SHAP más la estimación del puntaje medio. Podemos observar lo que ocurre en dos casos típicos de puntajes alto y bajo, ya reseñados en la sección 6, para observar como las variables pueden influir de forma diferenciada según el caso (figuras [25] y [26]):

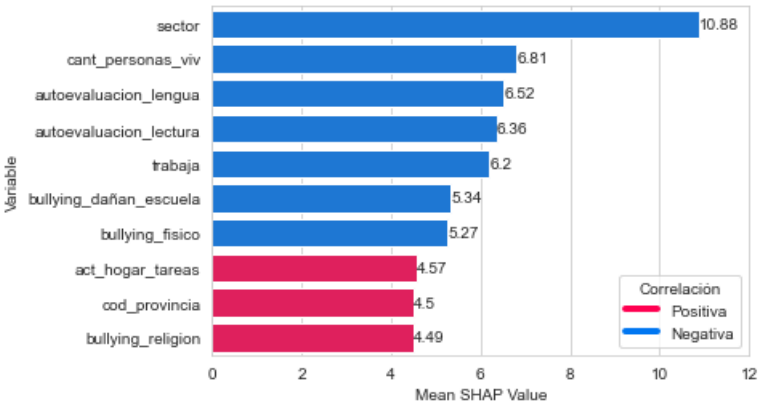


Figura [23]. Valores SHAP promedio por variable (mejores 10 variables).

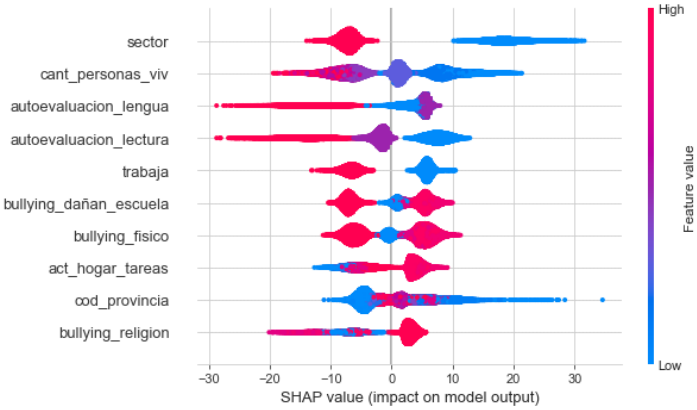


Figura [24]. Distribución de valores SHAP por variable según indicencia en variable objetivo.

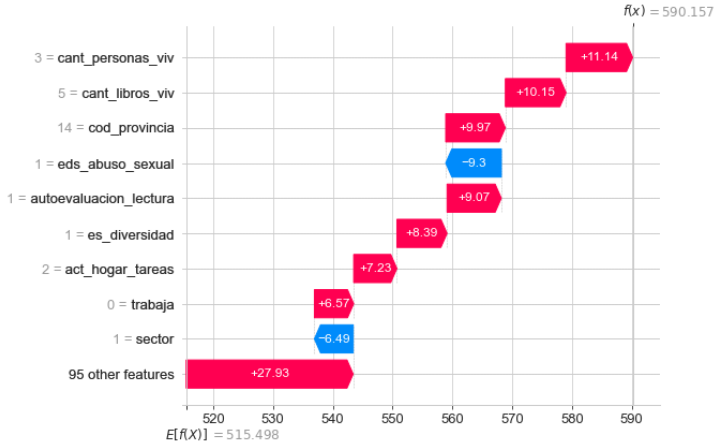


Figura [25]. Valores SHAP para alumno 197892.

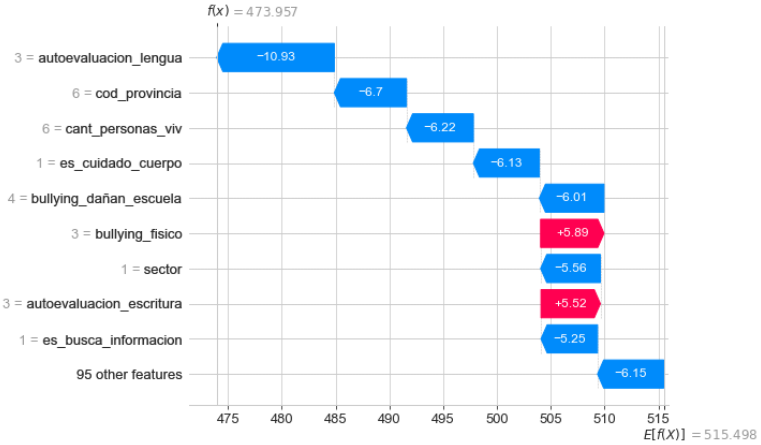


Figura [26]. Valores SHAP para alumno 293615.

- En el alumno 197892 (con un puntaje estimado de 75 puntos por encima de la media) están asociadas a un rendimiento positivo las variables cantidad de personas en el hogar (3), cantidad de libros (más de 100) y la provincia (Córdoba), entre otras, mientras que "reducen" el rendimiento las variables existencia educación sexual (sí) y la pertenencia al sector público.
- En cambio, observamos un comportamiento distinto en el alumno 293615, con un rendimiento predicho de casi 40 puntos debajo de la media. La autoevaluación en lengua relativamente mala (categoría 3, siendo 4 la peor) supone la mayor caída en puntaje, mientras que una autopercepción en escritura similar (autoevaluación_escritura=3) significa un incremento en el puntaje estimado. De la misma forma, la no existencia de bullying en la forma de daños a la escuela (categoría 4) reduce la estimación mientras que un nivel bajo de bullying físico (categoría 3) incrementa la estimación en similar cuantía.

8.2.3. SELECCIÓN DE VARIABLES (*FEATURE SELECTION*)

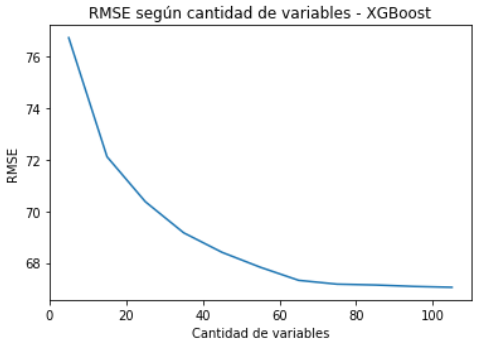


Figura [27]. RMSE según cantidad de variables - XGBoost

Realizamos un proceso de selección de variables (*feature selection*) a partir del modelo XGBoost con optimización bayesiana, tomando como referencia diferentes conjuntos de variables, a partir de la *feature importance* arrojada por el modelo (10 iteraciones con un paso de 10, entre las 5 y las 105 variables, ordenadas por su importancia, o aporte al modelo). Si bien no logramos mejorar el desempeño del modelo con todas las variables, podríamos -siguiendo el método del codo- seleccionar las mejores 65 variables. La caída en la performance es mínima (0,3 puntos de error) pero a cambio ganaríamos un modelo más robusto (esto es, más resistente al sobreajuste) a la vez que un menor costo en la adquisición de los datos (un cuestionario más acotado y facil de administrar).

8.3. AJUSTE FINAL

Realizamos un ajuste final del modelo con mejor desempeño -XGBoost con optimización bayesiana- sobre la totalidad del conjunto de entrenamiento, y obtuvimos los siguientes resultados:

RMSE CV	RMSE Test	R^2 Test	Tiempo de ejecución
66,8	66,7	0,41	00:36:00

Tabla [11]. Resultados en CV y conjunto de prueba (test) del modelo entrenado en conjunto de entrenamiento completo.

Podemos apreciar una leve mejoría en el desempeño en CV, con un RMSE de 66,8. La mejoría sobre el modelo de línea de base se mantiene en un 23%. El error en el conjunto de prueba es prácticamente igual. Como dato adicional introducimos la métrica R^2, o coeficiente de determinación. R^2 nos indica el porcentaje de la varianza explicada por el modelo, es decir, surge de dividir la suma de cuadrados de la regresión (diferencias entre el valor medio y la recta de la regresión) y la suma de cuadrados total (diferencias entre el valor observado y el valor medio):

$$R^2 = \frac{SS_{Regresión}}{SS_{Total}}$$

donde 0 nos indica un modelo que no explica nada de la variación de la respuesta alrededor de la media (y por lo tanto no es mejor que si tomaramos la media de referencia), mientras que 1 nos indica un modelo que explica toda la variación de la respuesta. Es decir, podemos decir que nuestro modelo explica el 41% de la varianza en la variable a predecir “puntaje de lengua”.

9. MODELO PARA PUNTAJE DE MATEMÁTICA

Replicamos el esquema de experimentos obtenidos anteriormente, de forma simplificada, para predecir el puntaje de la evaluación de matemática. Es decir:

1. Ajustamos un modelo XGBoost a la muestra de 80.000 casos.
2. Buscamos los mejores hiperparámetros a través de una optimización bayesiana.
3. Ajustamos el modelo final con la base completa.

A continuación se detallan los resultados de los experimentos:

Algoritmo	RMSE promedio en CV	Tiempo de ejecución	Hiperparámetros
Baseline	101,6	00:00:02	-
XGBoost*	84,1	00:01:33	Por defecto
XGBoost*	81,6	00:03:41	Optimización bayesiana
XGBoost*	82,2	00:03:20	Optimización bayesiana con selección de variables (n=55)

Tabla [12] Desempeño de modelos baseline vs XGBoost (matemática).

Podemos apreciar, a priori, mayores errores que respecto al modelo anterior. Sin embargo, vemos también reducciones en el error (contra el modelo baseline) similares a las encontradas anteriormente (alrededor del 20%). En lo que respecta a la selección de variables, vemos que un buen punto de equilibrio se encuentra en la utilización de 55 variables (ver figura [28]).

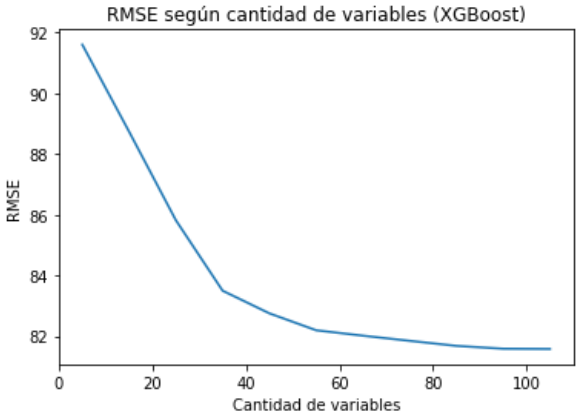


Figura [28]. RMSE según cantidad de variables – XGBoost (matemática)

RMSE CV	RMSE Test	R^2 Test	Tiempo de ejecución
80,8	80,8	0,37	00:34:15

Tabla [12]. Resultados modelo XGBoost CV y Test (matemática).

En lo que hace a la importancia de las variables, podemos apreciar en la tabla [13] que la variable más importante continúa siendo el sector de gestión, al igual que en el modelo de lengua, aunque en este caso con una importancia relativa bastante menor. Otras variables que se repiten con respecto al modelo anterior son la autoevaluación en el área en cuestión, la existencia de computadoras en el hogar, el hecho de trabajar y 3 variantes de bullying. Es llamativa la no aparición del índice socioeconómico del alumno y la aparición, en cambio, del ámbito de la escuela. Es notoria la aparición de la variable "cantidad de libros en el hogar", quizás como *proxy* del nivel cultural en el hogar. Por otro lado, las 10 variables más importantes parecen explicar un menor porcentaje del modelo respecto al modelo de lengua (33% vs 47%).

Variable	**Pregunta de referencia**	**Feature importance**
sector	Sector	0,07
nivel_autoev_m	Nivel autoevaluación en matemática	0,05
carac_viv_pc	¿Cuáles de estas cosas hay en el lugar donde vivís? Computadora	0,04
bullying_fisico	Por favor indicá cuántas veces los estudiantes de tu escuela…Discriminan por los aspectos físicos	0,03
bullying_nacionalidad	Por favor indicá cuántas veces los estudiantes de tu escuela…Discriminan por la nacionalidad	0,03
bullying_dañan_escuela	Por favor indicá cuántas veces los estudiantes de tu escuela…Dañan las cosas de la escuela	0,03
cant_libros_viv	Aproximadamente, ¿cuántos libros hay donde vivís?	0,02
ambito	Ámbito	0,02
trabaja	Trabaja (ya sea en el marco de la familia o no)	0,02
trabaja_con_familia	Además de asistir a la escuela, ¿ayudás a tus padres o familiares en su trabajo?	0,02

Tabla [13]. Importancia relativa de variables modelo XGBoost matemática (optimizado). 10 variables más importantes. Valores normalizados (sumatoria = 1).

Como cabe esperar, los resultados del modelo ajustado con el conjunto de entrenamiento completo arrojan resultados ligeramente mejores a los resultados a partir de la muestra, con resultados indistinguibles entre validación cruzada y *test*. Por otro lado, el coeficiente R2 nos indica que el modelo explica el 37% de la variablidad de los datos, algo menos que en el modelo de puntaje de lengua (41%).

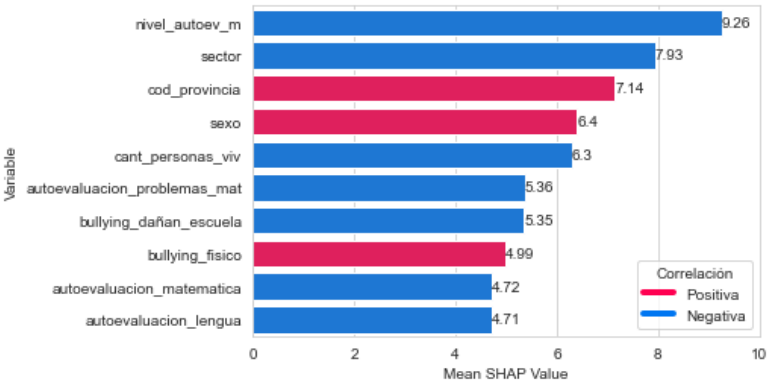


Figura [29]. Variables ordenadas por valores SHAP promedio – modelo para puntajes de matemática.

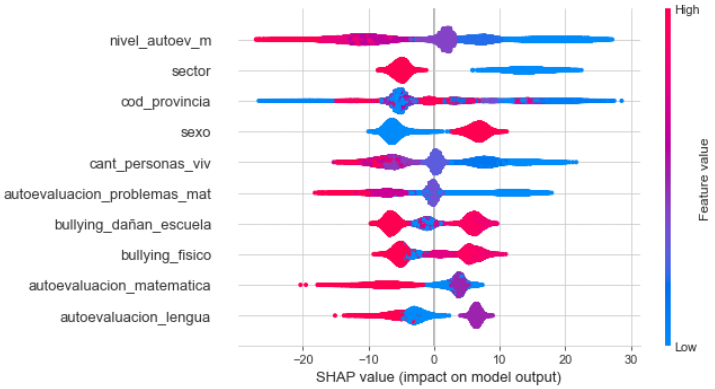


Figura [30]. Distribución de valores SHAP por variable según indicencia en variable objetivo – modelo para puntajes de matemática.

Observando los valores SHAP, nuevamente tenemos diferencias con respecto al listado original de importancia de variables arrojado por el modelo.

- Aparecen 3 variables asociadas a la autopercepción en matemática, y una asociada a la autopercepción en lengua. De la misma forma que en el modelo de puntajes de lengua, la correlación en realidad debe intepretarse como positiva, en tanto el sentido de la codificación está invertido (1 = muy buena autopercepción y 4= mala autopercepción). Es decir, a mejor autopercepción en matemática y lengua, peor desempeño, y viceversa.
- También vuelven a aparecer las variables provincia y cantidad de personas en el hogar.
- Las variables asociadas al bullying (daños a la escuela y religión) tienen un comportamiento no lineal (como en el modelo de lengua), con una mayor ocurrencia asociados levemente a una caída en el desempeño. Sin embargo, la correlación de la variables, en promedio, es *positiva* en el caso del bullying asociado a los daños en la escuela, y *negativa* en el caso del bullying físico (recordar que debe invertirse la interpretación para estas variables).
- Por otro lado, es notoria la aparición de la variable sexo (teniendo los varones un mejor desempeño esperado que las mujeres).

Como desarrollo final, dejamos disponible una aplicación online, a modo de ejercicio, con una versión simplificada de los modelos de lengua y matemática (tomando en cuenta solo las mejores variables), para obtener la predicción del modelo y correspondientes valores SHAP, a partir de los valores ingresados. Esta aplicación puede accederse a través de esta URL: http://aprender-predictor.herokuapp.com/.

10. CONCLUSIONES

Los modelos ajustados muestran un poder predictivo moderado, con una mejoría sobre el nivel basal del 23% y un coeficiente de determinación de 41%, para el caso de lengua, y del 19% y 37%, respectivamente, para el caso de matemática. Estos valores son indicativos de la complejidad del problema, y la dificultad de hacer predicciones con un alto grado de precisión, en base al nutrido cuestionario existente. Sin embargo, apreciamos como a partir de la administración de un cuestionario complementario limitado, con aproximadamente el 60% de las preguntas originales, es posible predecir con cierto nivel de precisión el puntaje de lengua esperado, lo cual puede ser útil en situaciones en que por cuestiones de recursos, tiempo o accesibilidad no sea posible administrar el cuestionario complementario completo, junto con las evaluaciones de lengua y matemática propiamente dichas.

Por otro lado, el análisis exploratorio realizado y los modelos predictivos entrenados permiten extraer algunas conclusiones valiosas sobre el desempeño de los estudiantes de 6to grado en lo que respecta a la prueba estandarizada de lengua (que pueden ser extendidas en gran parte a los modelos predictivos de matemática):

- Los 3 modelos analizados se muestran de acuerdo sobre la importancia de la variable “sector de gestión” al momento de estimar el puntaje de lengua. Desafortunadamente, vemos a las claras la desventaja de pertenecer a escuelas de gestión pública en comparación con las escuelas de gestión privada. Una política de fuerte inversión y mejora de la calidad educativa en este sector se avisora como un paliativo urgente y necesario.

- El hecho de que el niño trabaje también aparece como una variable importante en el modelo de regresión y en el modelo XGBoost. Es claro que el hecho de realizar actividades laborales a temprana edad significa un perjuicio en la calidad educativa que el niño puede recibir. Además, en el caso de los niños que trabajan fuera del ámbito familiar, constituye una violación flagrante a la ley 26.390 de la Constitución Nacional, que prohibe que los menores de 16 años trabajen.

- El índice socioeconómico del alumno aparece como una variable importante en los modelos de árboles de decisión. Si bien podemos establecer una correlación entre esta variable y el sector de gestión de pertenencia, el índice socioecónomico es una variable que atraviesa al sector de gestión, en tanto podemos encontrar niños de nivel bajo que estudian bajo la órbita del sector privado, a la vez que niños de nivel alto en el sector público. Una política pública deseable en este sentido sería el apoyo económico específico para estos sectores postergados.

- Las variables asociadas a las diversas modalidades de bullying parecen ser significativas en los modelos delengua y matemática. Es de esperar que un trabajo en las aulas que trate esta problemática ayudaría a mejorar las condiciones de la enseñanza y el aprendizaje, en pos de lograr mejores resultados académicos.

- La variable provincia se reveló como significativa al computar los valores SHAP, tanto para matemática como para lengua. Esto revela las profundas diferencias en la calidad educativa de las distintas regiones del país.

Sin dudas, son muchas las dimensiones en las cuales el sistema educativo en general, y el sistema educativo público en particular, pueden realizar mejoras sobre los sectores sociales más perjudicados. Esperamos que este trabajo haya realizado aportes en este sentido, y pueda ser complementado con análisis de mayor profundidad en el futuro.

11. BIBLIOGRAFÍA

[1]. Chen, T, Guestrin, C, "XGBoost: A Scalable Tree Boosting System" (https://arxiv.org/pdf/1603.02754.pdf), 2016

[2]. Cortez, P., Silva, A., "Using data mining to predict secondary school student performance", Dep. Information Systems/Algoritmic R&D Centre, University of Minho, 2008

[3]. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

[4]. https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

[5]. Szretter Noste, M. E., "Apunte de regresión lineal", Carrera de Especialización en Estadística para Ciencia de la Salud, FCEN UBA, 2017

[6]. Tibshiriani, R., Witten, D., Hastie, T., James, G., "An Introduction to Statistical Learning" (traducción propia), Springer, 2013

[7]. "APRENDER 2018 - INFORME NACIONAL DE RESULTADOS - 6°AÑO NIVEL PRIMARIO", Ministerio de Educación, Cultura, Ciencia y Tecnología de la Nación Argentina, 2018

[8]. "EVALUACIÓN NACIONAL APRENDER - DOCUMENTO METODOLÓGICO", Ministerio de Educación, Cultura, Ciencia y Tecnología de la Nación Argentina, 2018

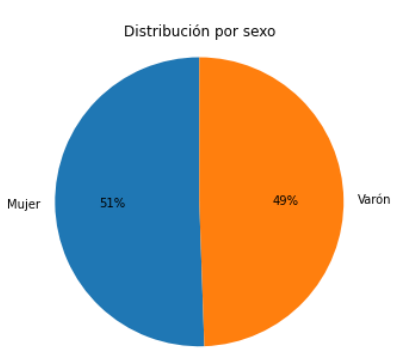
[9]. Base de datos completa disponible en https://drive.google.com/file/d/18Y0gy0ZsjT4n7e9hlSOtCuvI0DFdhJt6/view/Base estudiantes 6 grado primaria 2018 USUARIA.csv

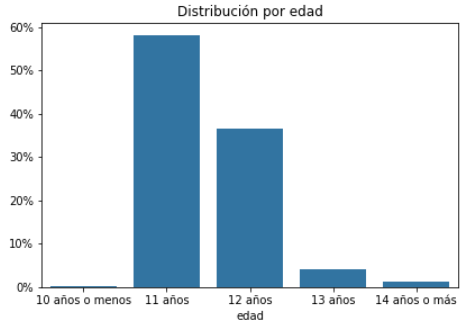
[10]. Chan, D., Badano, C. Rey, A., "Análisis inteligente de datos con lenguaje R", Apunte para Maestría en Explotación de Datos y Descubrimiento del Conocimiento, 2018

[11]. Pagès, J., "Multiple Factor Analysis by Example Using R", CRC Press, 2015

12. ANEXO METODOLÓGICO

Algunas variables adicionales del análisis exploratorio:





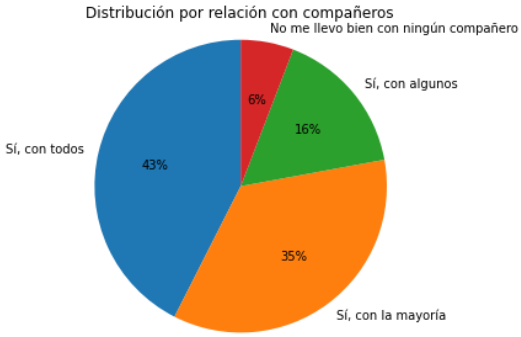


Figura [31] (Izquierda). La distribución en términos de sexo es casi completamente equitativa, con una ligera diferencia en favor de las niñas.
Figura [32] (Medio). La gran mayoría de los estudiantes tienen 11 o 12 años, lo que es esperable por tratarse de estudiantes de 6to grado. Cabe esperar que los estudiantes de mayor edad sean repitentes o correspondan a escuelas rurales.
Figura [33] (Derecha). Hay una distribución bastante pareja de las categorías positivas de relación con los compañeros (con todos, la mayoría o algunos). Un 1% de los estudiantes no se lleva bien con ningún compañero.

Posibles respuestas asociadas a preguntas involucradas en ingeniería de variables:

Valor numérico	Etiqueta
1	Siempre
2	Muchas veces
3	Pocas veces
4	Nunca

Tabla [14]. Variables act_hogar_cuidado, act_hogar_tareas, act_hogar_cultivo.

Valor numérico	Etiqueta
1	Siempre
2	La mayoría de las veces
3	Algunas veces
4	Nunca

Tabla [15]. Variables autoevaluacion_lengua, autoevaluacion_matemática.

Valor numérico	Etiqueta
1	Muy bien
2	Bien
3	Más o menos bien
4	No muy bien

Tabla [16]. Variables autoevaluacion_lectura, autoevaluacion_escritura, autoevaluacion_matematica.

Si bien las categorías de las variables ap20 y ap22, por un lado, y ap21a, ap21b y ap23 por el otro, no son las mismas, se decidió combinarlas en tanto el número de categorías es el mismo.

Variable	Nombre variable	Categorías recodificación
Cantidad de habitaciones	cant_hab	Bajo/Medio/Alto
Nivel características vivienda	nivel_carac_viv	Bajo/Alto
Nivel de actividad en el hogar	nivel_act_hogar	Bajo/Alto
Nivel de bullying	nivel_bullying	Bajo/Alto
Nivel de actividades extraexcolares	nivel_actividades	Bajo/Medio/Alto
Nivel de educación sexual	nivel_eds	Bajo/Medio/Alto
Nivel de autopercepción en lengua	nivel_autoev_l	Bajo/Alto
Nivel de autopercepción en matemática	nivel_autoev_m	Bajo/Medio/Alto

Tabla [17]. Recodificación de variables numericas en categóricas para MCA.

Puntaje de lengua según niveles:

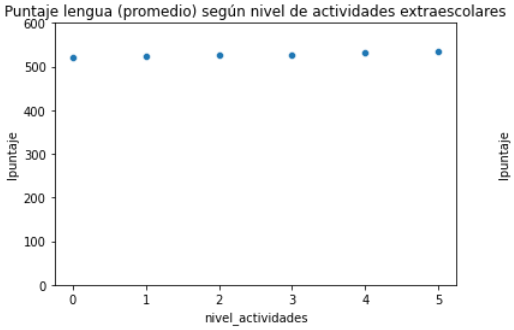


Figura [34]. Puntaje lengua según nivel de act. extra-escolares

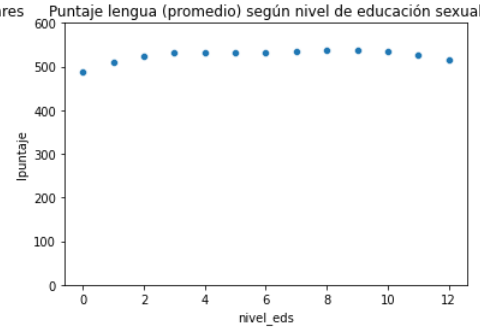


Figura [35]. Puntaje lengua según nivel de educación sexual

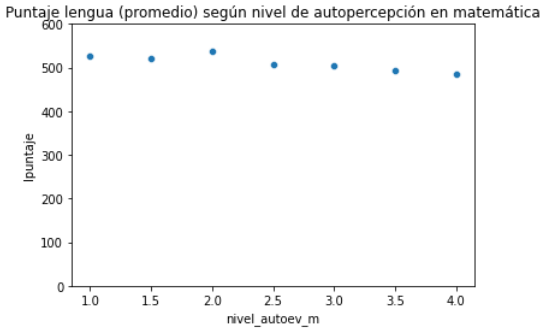


Figura [36]. Puntaje lengua según nivel de autopercepción en matemática.

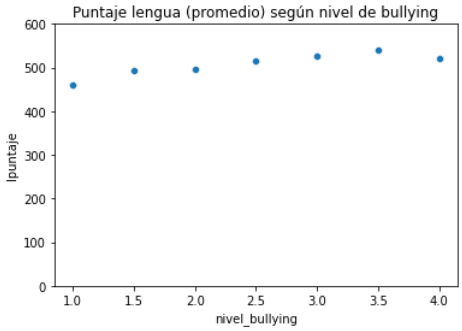


Figura [37]. Puntaje lengua según nivel de bullying

Variable renombrada	Target	Predictor	Modelo final FS Lengua	Modelo final FS Matemática
edad		X		
sexo		X		X
pais		X		
pais_mama		X		
pais_papa		X		
cant_personas_viv		X	X	X
viv_madre		X		
viv_padre		X		
viv_tutor		X		
viv_hermano		X	X	X
viv_tio		X	X	X
viv_abuelo		X		
viv_otro		X	X	X
viv_desc		X		
cant_hab		X		X
carac_viv_internet		X		X
carac_viv_agua		X	X	X
carac_viv_pc		X	X	X
carac_viv_heladera		X		
carac_viv_aa		X		
carac_viv_calefaccion		X		
cant_libros_viv		X	X	X
max_nivel_ed_mama		X	X	X
max_nivel_ed_papa		X		X
fam_indigena		X	X	X
lengua_indigena		X	X	
act_hogar_cuidado		X	X	X
act_hogar_tareas		X	X	X
act_hogar_cultivo		X	X	X
trabaja_con_familia		X	X	X
trabaja_sin_familia		X	X	X
asistencia_jardin		X	X	X
repitio		X	X	X
gusta_escuela		X		
buena_relacion_comp		X	X	X
autoevaluacion_lengua		X	X	X
autoevaluacion_lectura		X	X	X
autoevaluacion_escritura		X	X	X
autoevaluacion_matematica		X	X	X
autoevaluacion_problemas_mat		X		X
maestros_vuelven_explicar		X	X	X
bullying_buenas_notas		X	X	X
bullying_repitieron		X	X	X
bullying_religion		X	X	X
bullying_fisico		X	X	X
bullying_discapacidad		X	X	
bullying_nacionalidad		X		X
bullying_dañan_escuela		X	X	X
bullying_violencia_fisica		X	X	

actividades_deporte		X	X	
actividades_lectura		X	X	X
actividades_amigos		X		X
actividades_idioma		X		X
actividades_espectaculo		X	X	X
celular		X		
celular_internet		X		
es_cambios_cuerpo		X	X	
es_cuidado_cuerpo		X	X	X
es_embarazo		X		
es_ets		X	X	
es_derechos		X	X	
es_igualdad_genero		X	X	
es_diversidad		X	X	
es_prevencion		X	X	
es_abuso_sexual		X	X	
es_pedir_ayuda		X	X	
es_comunicacion		X		
es_buen_trato		X		
es_com_escuela		X	X	X
es_com_escuela_otros		X	X	
es_com_amigos		X		
es_com_fam_mujer		X		
es_com_fam_varon		X		
es_com_otros		X		
es_com_nadie		X	X	
es_cambios_cuerpo		X	X	
es_cuidado_cuerpo		X	X	X
es_embarazo		X		
es_ets		X	X	
es_derechos		X	X	
es_igualdad_genero		X	X	
es_diversidad		X	X	
es_prevencion		X	X	
es_abuso_sexual		X	X	
es_pedir_ayuda		X	X	
es_no_interesa		X	X	X
es_busca_informacion		X	X	X
transporte_colectivo				
transporte_auto				

transporte_bici				
transporte_pie				
transporte_caballo				
transporte_otro				
tiempo_viaje				
act_esc_establecimientos				
act_esc_excursiones_ciudad				
act_esc_deportes_cultura				
trabajos_en_grupo				
ayuda_pequeños				
trabajo_grupal				
tareas_dificiles				
responden_grandes				
falta_por_trabajo				
falta_por_trabajo_dias				
prospecto_secundaria				
cercania_secundaria				
dificultad_lengua				
dificultad_matematica				
dificultad_cuestionario				
dificultad_vision				
cod_provincia		X	X	X
sector		X	X	X
ambito		X	X	X
ICSE		X		
ponder				
lpondera				
ldesemp				
lpuntaje	X			
mpondera				
mdesemp				
mpuntaje	X			
isocioa		X	X	X
isocioa_puntaje		X		
isocioal				
isocioam				
migración		X		
hogar_indigena		X		
trabaja		X	X	X
max_nivel_socio_ed_padres		X		
nivel_carac_viv		X	X	
nivel_act_hogar		X		
nivel_bullying		X		
nivel_actividades		X		X
nivel_eds		X	X	X
nivel_autoev_l		X		
nivel_autoev_m		X	X	X

Tabla [18]. Listado de variables completo, indicando variables utilizadas en modelo de lengua y matemática con selección de variables (FS).