Educación Secundaria en Argentina: Factores Condicionantes al Desempeño Académico

Alumno: Matias Oscar Grouman

Director: Gastón Pezzuchi

Co Directores: Ana Laguna Pradas

Juan Ale

Facultad de Ingeniería – Universidad Austral

Buenos Aires, Argentina 2021

Índice General

Índice de Gráficos

Índice de Tablas

Resumen

En los últimos años la tecnología generó cambios de paradigma, cuestionando el contexto en el que vivimos. La educación es un elemento imprescindible para que la sociedad pueda desarrollarse. Por lo tanto, el sistema educativo debe evolucionar a la par de las necesidades de sus usuarios. En la Argentina la educación no satisface los requerimientos de los alumnos y el escenario empeora cuando pensamos en el futuro, mientras los cambios de paradigma son exponenciales año tras año. Por este motivo, la educación necesita de información relevante para responder ante las necesidades de la sociedad. En consecuencia, uno de los objetivos del trabajo es estudiar el motivo por el que un alumno, del último año de secundaria, tiene mejor desempeño en las áreas de Matemática y Lengua, ya que son habilidades que tomarán relevancia a lo largo del tiempo para facilitar el dominio de las nuevas tecnologías. La investigación consiste en analizar los resultados del censo Aprender 2016, donde se relevó el desempeño de los alumnos y la forma en la que viven y aprenden, a través de un examen teórico y práctico para cada materia (Lengua, Matemática y Ciencias) y un cuestionario anexo. De esta manera se van a detectar variables que condicionan el desempeño académico permitiendo que los educadores las potencien y así mejorar la educación. Junto con el resultado final, se obtendrán recomendaciones sobre áreas para profundizar las próximas encuestas y así retroalimentar el modelo. Por último, a lo largo de la investigación se abordan temas de interés general, tales como la percepción de la tecnología, el trabajo en equipo y la relación del nivel socioeconómico y la educación.

Palabras claves

Aprender, Desempeño, Educación, Lengua, Matemática, Tecnología.

Capítulo 1: Introducción

1.1 Motivación e Importancia del Campo

Puesto que nos encontramos en un mundo globalizado y en constante cambio, brindar herramientas al sistema educativo, para que pueda seguir el ritmo de los mismos, resulta importante para asegurar un futuro mejor (Fuego Simondet, 2019). A su vez, el Gobierno de Nueva Zelanda, afirma en su estudio sobre la Educación del futuro que el contexto mundial es complejo y el aprendizaje debe convivir con el constante cambio de las habilidades requeridas en el mundo laboral, gran cantidad de información disponible y alta adaptación a los cambios (New Zealand Government, 2015). Mas aún, los establecimientos educativos se enfrentan a variados desafíos, tales como la nueva identidad de los docentes con la presencia de tecnología y las necesidades del mundo actual (Garabito & Vezub, 2017).

Como nadie ignora, los niños que hoy están en la escuela son aquellos que sostendrán la economía del país en los próximos años (Rivas, 2010). Y la herramienta principal para formarse ante los desafíos que se les vayan a presentar, es la educación. Por lo tanto, toda información relevante sobre el área tiene un impacto social (Fundación Universidad Católica, 2020).

A pesar que la educación es uno de los Derechos Humanos (Naciones Unidas, 2020), la cuestión es saber si el acceso es suficiente o si los Gobiernos deben asegurar otras herramientas para que los niños aprendan bajo las mismas condiciones. Este trabajo brinda información sobre la existencia de factores que aseguren las mismas oportunidades de aprendizaje. De esta manera, se va a mejorar la educación lo que contribuye a una sociedad libre, según Horace Mann (Broome, 2018).

A partir de 1974, donde el 32% de las personas terminaba el secundario, se ha observado una tendencia positiva, duplicando el valor al cierre del año 2018. Este porcentaje se calculó tomando una muestra de individuos entre 18 y 29 años de la Provincia de Buenos Aires (Fundación Universidad Católica, 2020). Con este acercamiento, podemos afirmar que el acceso a la educación ha tenido mejoras a lo largo del tiempo, ahora es momento que la información con la que cuentan las Instituciones permita contribuir al mejor desempeño de sus alumnos.

Dado que, los recursos son limitados, el tiempo corre y los avances tecnológicos más aún, el foco de este análisis es brindar herramientas que contribuyan al desarrollo educativo y contribuir con la enseñanza del país. Así las Instituciones y el Gobierno pueden ser más eficientes.

En la conclusión del trabajo, se puede determinar qué ajustes son necesarios en la educación o sociedad, para que un alumno pueda aprender bajo las mismas condiciones, asegurando igualdad de oportunidades. También se obtienen resultados sobre la presencia de la tecnología en las aulas, y el impacto del nivel socioeconómico en la educación.

1.2 Requerimientos y Desafíos

La Educación cuenta con variados análisis provenientes del Data Mining, por un lado se ha trabajado en la segmentación (Clemens, Malbernat, Urrizaga & Varela, 2015), tanto en alumnos, para brindar educación personalizada, como en docentes, para agruparlos según sus capacidades sobre la digitalización de sus contenidos. Con respecto a la educación personalizada se han realizado pruebas piloto con algoritmos de inteligencia artificial que muestran la posibilidad de aumentar hasta un 15% el éxito en los exámenes escolares (Banco Interamericano de Desarrollo, 2018).

Por otro lado, se realizaron trabajos predictivos para el desempeño de los alumnos o para sugerir áreas de estudio universitario. Para ilustrar, un trabajo realizado sobre la predicción en las notas de alumnos universitarios fue a través de la aplicación de árboles de decisión, con el objetivo de clasificar a los alumnos en base a los datos almacenados, tales como asistencia, si trabaja en investigaciones y puntajes de otros cursos. (Baradwaj & Pal, 2011).

Específicamente sobre la encuesta Aprender, los trabajos que se encuentran son más que nada descriptivos. El Estado brinda una herramienta para visualizar los datos obtenidos y analizarlos, principalmente con gráficos y destinado a docentes. Por otro lado, en el trabajo de Ivana Templado de la Fundación de Investigaciones Económicas Latinoamericanas (2018) se investiga el desempeño de los alumnos de 2016 y se aborda el papel de la tecnología en la investigación, tomando como base los resultados de Aprender de dicho año. A su vez, aquí se han tomado algunas variables para relacionarlas con el desempeño de los alumnos, utilizando *tests* de hipótesis.

Dado que, el pensamiento lógico y la interpretación de textos han sido definidos como dos aspectos del ser humano vitales para convivir con las máquinas, las materias más relevantes para desarrollar estas cualidades, son Matemática y Lengua (Fuego Simondet, 2019). En 2030 se espera que los trabajadores usen un 41% más de pensamiento crítico y razonamiento, tras la inclusión de las máquinas en sus puestos de trabajo (Banco Interamericano de Desarrollo, 2018). En efecto, la principal utilidad

del trabajo es saber por qué un alumno tiene mejor puntaje en estas áreas que resultan fundamentales para el futuro. A su vez, conocer la percepción de los alumnos sobre el uso de la tecnología en las aulas y la utilidad que le brindan fuera de ellas.

Para resumir, estas conclusiones generan múltiples beneficios, por un lado, en la toma de decisiones sobre la educación y políticas sociales, pero también es una herramienta para los educadores donde pueden saber con certeza qué factores incentivan el aprendizaje de sus alumnos. Por último, surgen recomendaciones sobre áreas para profundizar próximas encuestas y así retroalimentar el trabajo y obtener mayor información.

1.3 Problemas no Resueltos

Los egresados de las escuelas secundarias no tienen las herramientas adecuadas para el contexto en el que viven. Esto sucede por variados motivos, pero uno de ellos es la falta de información relevante sobre la situación actual, donde se describan las necesidades específicas de los alumnos y se puedan tomar decisiones al respecto. Este trabajo contribuirá con resultados contundentes sobre los puntos a mejorar en el ámbito académico y así contribuir a una mejor sociedad.

Tal como se ha mencionado en párrafos anteriores, todo lo establecido es cuestionado, donde nos encontramos con afirmaciones como la de Alvin Toffer donde sostiene que el analfabeto del futuro no será la persona que no pueda leer, sino aquel que no sepa cómo aprender (Rivas, 2010). En consecuencia, nos preguntamos cómo aprenden las personas, qué los incentiva y en qué medida. El presente trabajo apunta a responder esas preguntas, identificando los factores determinantes al aprendizaje para así poder mejorarlo.

Puesto que la Educación forma personas para emprender sus proyectos y desarrollarse en la sociedad, todo aporte para mejorarla tiene impacto social, más aún en el contexto que nos encontramos donde los cambios de paradigma se dan cada vez más rápido y con mayor magnitud.

En consecuencia, medir datos educativos es relevante para la sociedad ya que permite conocer las capacidades de los alumnos escolares, quienes serán la población económicamente activa en unos años. A su vez da lugar a comparar las distintas regiones de la sociedad para encontrar oportunidades (Rivas, 2015). En otras palabras, la “datificación” es quizás la fuerza más poderosa del nuevo mundo educativo digital ya que permite medir aspectos que nunca antes se han medido por los centros de control

de la educación. A lo largo del tiempo se han tomado medidas sobre metodologías de educación, por ejemplo en 2017 Francia lanzó una regulación de tolerancia cero en las escuelas y los celulares (Rivas, 2010). Este es otro punto que se responderá en la investigación, favoreciendo al Estado e Instituciones a la hora de tomar este tipo de decisiones.

1.4 Contribución del Trabajo

La solución que se obtiene al finalizar el trabajo es el grado de relación entre las preguntas del cuestionario anexo y el desempeño de los alumnos, entendiendo por qué un alumno de secundaria es mejor que otro en Matemática y Lengua.

Además, el análisis trae soluciones aparejadas, tales como le percepción que tienen los alumnos sobre la tecnología en el aula y el uso de sus tiempos libres. Otro aspecto abordado es la relación del índice socioeconómico y el grado de estudios de los padres con su desempeño.

Por último, se busca retroalimentar el censo que actualmente realiza el Estado, sugiriendo áreas para hacer foco en las próximas encuestas, aumentando los beneficios resultantes del modelo que abordaremos.

1.5 Estructura del Trabajo

Para alcanzar los resultados establecidos, se trabajó sobre la base de datos del año 2016. Luego se aplicaron algoritmos de regresión y *feature importance*, con el fin de identificar las variables de mayor impacto en el *target*.

Por otro lado, se hicieron sugerencias sobre campos o áreas para profundizar en las próximas encuestas, y así enriquecer el primer análisis con la posibilidad de ver tendencias.

Las encuestas Aprender fueron un nuevo proyecto en Argentina, que aún está amoldándose, inclusive el Gobierno ya ha anunciado un nuevo relevamiento de los estudiantes para conocer cómo fue el impacto del COVID en su aprendizaje, en otras palabras, una evolución de Aprender. Entre Aprender 2016 y 2017 se observan diferentes formas de preguntar lo mismo, dificultando ver evoluciones. También al tratarse de niños hay datos sobre sus padres o sus hogares que no saben, dando lugar a valores faltantes o erróneos.

1.6 Objetivos del Trabajo

Objetivo General:

Analizar el desempeño relevado en Aprender 2016 de los alumnos de secundaria en las materias Matemática y Lengua, explicando qué factores afectan el mismo.

Objetivos Específicos:

1) Vincular el índice socioeconómico y su influencia en el desempeño académico.
2) Determinar si la tecnología tiene un impacto positivo en la educación.
3) Brindar campos de estudio para las próximas encuestas.
4) Determinar el papel de los padres en la educación de sus hijos.

1.7 Transferencia de los Resultados Obtenidos

El análisis podría replicarse en otros años de estudio, dentro de la secundaria o incluso en la primaria. En el futuro, podría aplicarse en las Universidades.

También se podría crear un ID anónimo para cada alumno y así tener un seguimiento de su desempeño y las variables que se estudien.

Resolver el problema sobre el acceso a la educación y los elementos que la condicionan, permitirá a las escuelas focalizarse en analizar qué necesitan modificar o incorporar para brindar herramientas que sean de utilidad en el futuro.

Es decir, el trabajo resuelve la cuestión de elementos para la educación. Lo que sigue es trabajar sobre las metodologías de aplicación.

Capítulo 2: Trabajos Relacionados

En la Argentina hay trabajos realizados sobre la educación, donde se afirma que el nivel socioeconómico es un factor relevante al desempeño de los alumnos, sosteniendo que el desempeño de los alumnos es afectado por los recursos de las familias y el contexto en el que viven. Aquí se analizó la relación entre el *status* social y el abandono de la escuela, donde se observaba un mayor porcentaje de abandono cuando se bajaba en la pirámide social, por ejemplo, algunos alumnos de bajos ingresos tendían a abandonar la escuela por la necesidad de trabajar para sobrevivir. El 33% de los hogares con menor ingreso per cápita familiar (primer tercil) concentra el 39% de los jóvenes entre 12 y 17 años que abandonaron la escuela (Fundación Universidad Católica, 2020).

Como es por muchos conocido, la encuesta Aprender surge de un estudio que comenzó en el año 2000 por parte de un grupo denominado PISA, perteneciente a la Organización para la Cooperación y el Desarrollo Económico (OCDE). Cuyo objetivo es medir el desempeño de los alumnos de quince años en la aplicación de matemática, comprensión lectora y ciencias. En este caso si hay mayor cantidad de trabajos realizados (Rivas, 2015).

Por su parte, el concepto "*Educational Data Mining*" (EDM) hace referencia a las investigaciones que buscan entender el aprendizaje a través de herramientas de Data Mining (Prabha Lakshmi, 2014). En esta investigación se aborda un caso de aplicación sobre la evolución del modelo de aprendizaje de cada alumno, de esta manera se podría medir el impacto de los cambios de metodologías.

A su vez, en La Enciclopedia de la Educación (tercera edición), Ryan S.J.d. Baker enumera las técnicas de Data Mining con su posible aplicación práctica. Particularmente en el área de Relacionamiento, lo que se abordó en esta investigación, hace mención sobre la búsqueda de relaciones entre una variable *target* y un conjunto de variables. Por ejemplo, si un alumno está frustrado y quiere aprender, entonces pedirá ayuda. Luego el *paper* analiza aplicaciones predictivas, como podría ser la predicción de que un alumno acierte o se equivoque en determinada tarea (Baker, 1989).

Y con respecto a la Educación Universitaria, se han hecho trabajos de regresión logística, cuyo objetivo fue identificar el motivo por el que un alumno abandona la carrera antes de finalizarla. Este análisis también encontró una relación entre el *status* económico y la deserción educativa, el 55% de los desertores se encuentra en el nivel más bajo de la pirámide social (Adrogue & Garcia, 2015).

Como se puede ver en los párrafos anteriores, aplicaciones de Data Mining en Educación ha habido muchas a lo largo de los años y en distintas partes del mundo, tales como análisis para dar *feedback* constructivo a los alumnos con herramientas de *clustering* y clasificación, hasta recomendaciones de aprendizaje virtual basado en el hábito de navegación de los alumnos. También fue utilizado para detectar el modelo de aprendizaje de los alumnos, viendo qué los motiva, donde predominan herramientas de *clustering*. Por último, relacionado a este trabajo, es la agrupación de estudiantes basado en características o variables determinadas (Romero & Ventura, 2010). En este último se encuentran trabajos supervisados (clasificación) y no supervisados (*clustering*), hasta el momento no se han encontrado análisis de este tipo para las encuestas Aprender, por lo tanto, se espera obtener información relevante para el Estado y las Instituciones.

Capítulo 3: Metodología

Se trabaja sobre la base de datos compartida por el Ministerio de Modernización, correspondiente a Aprender 2016. Con respecto al tipo de investigación, se trata de un estudio explicativo ya que busca establecer una relación de tipo causa – efecto entre la variable objetivo y el cuestionario anexo (Babin, Carr, Griffin & Zikmund, 2012).

En primer lugar, se exploran los datos, conociendo el mínimo nivel de expresión y así poder determinar la forma en la que se abordaran los objetivos. Esto es llevado a cabo a través de la herramienta Jupyter Python.

El cuestionario abarca variables sobre los alumnos, sus familias, lugar de residencia y hábitos. En principio no se excluye ninguna variable, pero a lo largo de la investigación podrían eliminarse variables redundantes. Y no se descarta la posibilidad de agregar o crear variables que contribuyan al modelo.

Al finalizar el proceso de exploración, se procederá con la búsqueda de modelos estadísticos para analizar la relación entre el *target* y las variables del cuestionario anexo.

3.1 Plan de Trabajo

A. Etapa 1: Exploración de los datos
    a. Exploración de las preguntas del cuestionario anexo y categorización de las mismas.
    b. Exploración de los datos a nivel general y sectorizado para comprender distribuciones y frecuencias.
B. Etapa 2: Transformación de los datos
    a. Definición de valores faltantes.
    b. Definición de valores *outliers*.
    c. Agregación de variables
C. Etapa 3: Exploración de los datos transformados
D. Etapa 4: Objetivos Generales
    a. Trabajar sobre la percepción de la tecnología y el índice socioeconómico.
E. Etapa 5: Modelado
    a. Identificar el modelo adecuado y ejecución para construir las conclusiones.

Capitulo 4: Solución Planteada

4.1 Preparación de los Datos

Aprender es una encuesta censal, donde se mide el desempeño de los alumnos en las materias de Matemática, Lengua, Ciencias Sociales y Ciencias Naturales. La medición es de alcance nacional y comprende más de 28 mil escuelas. Es ejecutada de forma estandarizada y en el mismo momento, buscando entregar información robusta y confiable sobre la educación. Se acompaña por un cuestionario anexo, donde se relevan variables sociodemográficas y referentes al estilo de vida del alumno, dentro y fuera de la escuela, con el objetivo de entender cómo viven y cuáles son sus oportunidades e intereses.

El *dataset* cuenta con un registro anónimo por alumno, con el resultado del examen para cada materia y la respuesta del cuestionario anexo. El total de registros es de 331.852. Producto de las diferencias entre las características propias a cada Provincia de Argentina, se consideró necesario tomar una región que mitigue esta heterogeneidad, pero sin perder gran cantidad de registros. En consecuencia, para el análisis se seleccionaron los datos correspondientes a la provincia Buenos Aires, incluyendo a la Ciudad Autónoma de Buenos Aires, lo que resulta en 137.570 registros (42% del *dataset*), con un 19% de *missings*. Dado que en Buenos Aires hay 159.155 individuos entre diecisiete y dieciocho años (Argentinos por la Educación, 2011), podemos afirmar que el *dataset* es representativo de la población.

En primer lugar, se buscó el nombre del municipio, ya que la base de Aprender solo contiene el código ID de Municipio. Esto es posible con la unión a la base del Instituto Geográfico Nacional (Datos Argentina, 2020), donde se identificó cada uno y se agregó una variable llamada "*Subcategory*" que segmentó los municipios en tres: Ciudad Autónoma de Buenos Aires (CABA), Área Metropolitana de Buenos Aires (AMBA) sin CABA e Interior de Gobierno de Buenos Aires (GBA). Esta última incluye todas las localidades de GBA que no están comprendidas dentro del AMBA, dichos valores se definieron según el Gobierno de la Ciudad de Buenos Aires (Gobierno de la Ciudad de Buenos Aires, 2020).

Por otro lado, las preguntas del cuestionario anexo fueron clasificadas según las Notas Metodológicas de Aprender (Aprender, 2020). De esta manera, se podrán identificar las variables referentes a Tecnología, Sociodemográfica, entre otras.

Dado que las variables son de opción múltiple, tienen un valor mínimo y un máximo, por lo tanto, con una sumarización de los datos se puede validar que no hay

presencia de valores nulos en el cuestionario anexo ya que no se observan registros fuera de los límites para cada pregunta. Por otro lado, las respuestas “no sabe”, “no contesta” (-1 y -9) se agruparon como valores *missings*, para ello se estableció un condicional donde si el valor es menor que cero se determina como valor perdido.

Con respecto a los valores faltantes, la base viene con un tratamiento con el módulo MI de Stata, donde se aplicaron regresiones logísticas para imputar los datos en base a las preguntas respondidas. En el caso donde el cuestionario está respondido en un 50% o menos, se descartó del *dataset* (Aprender, 2016).

Dado que el Nivel Socioeconómico será una de las variables más importantes en el trabajo, es importante mencionar cómo se construye el valor. Para lograr un indicador relevante, se corrieron dos procedimientos estadísticos: análisis de componentes principales y la teoría de respuesta al ítem. En base a variables del cuestionario se construyeron scores que resultarán en el nivel socioeconómico. Las variables utilizadas son:

- Nivel Educativo de los Padres.
- Hacinamiento del Hogar, resultado de la relación entre cantidad de habitaciones en el hogar y los miembros que viven en el mismo.
- Tenencia de Equipamiento Informático en el Hogar.

Para comenzar se hizo un entrenamiento con la base de la Encuesta Anual de Hogares Urbana que realiza el Instituto Nacional de Estadística y Censos (INDEC) y luego se aplicó el modelo a los datos de Aprender. Con el objetivo de validar el índice, se calculó una correlación de Spearman, con la variable objetivo “isocioa” que indica el nivel socio económico del alumno, se observa una correlación de 0.65 con el nivel educativo del padre y 0.55 con el de la madre. Tal como se puede observar en los siguientes dos gráficos, el nivel educativo de los padres presenta un mayor número de personas con estudios terciarios o universitarios en el AMBA que en el Interior. También hay diferencias entre los sexos ya que las Madres muestran un mayor grado de estudios que los Padres.

Gráfico 1: Nivel Educativo de la Madre (Ap7)

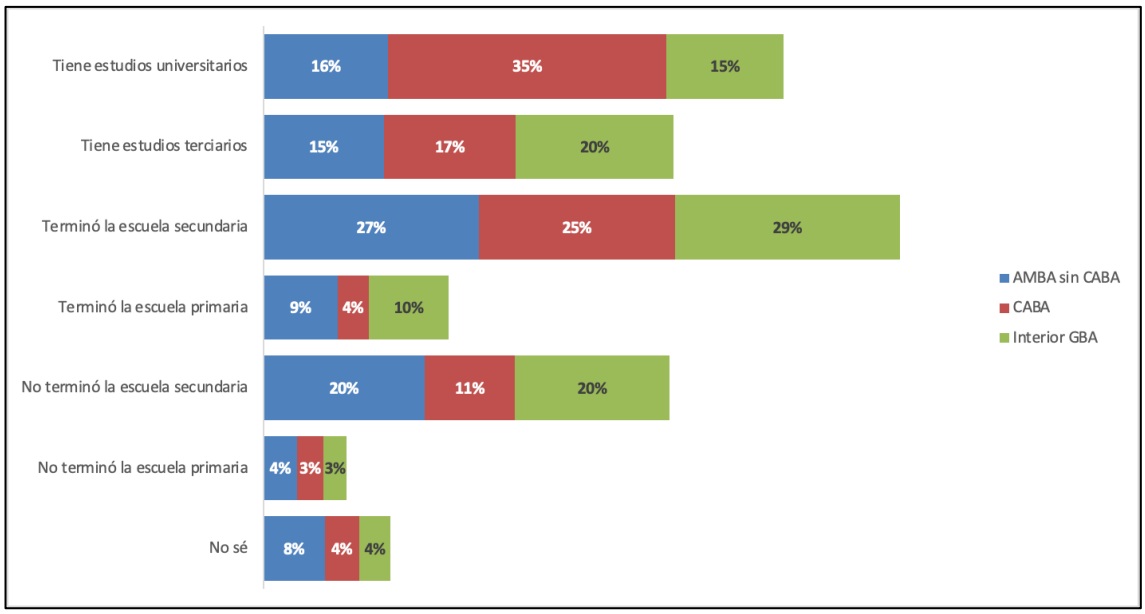
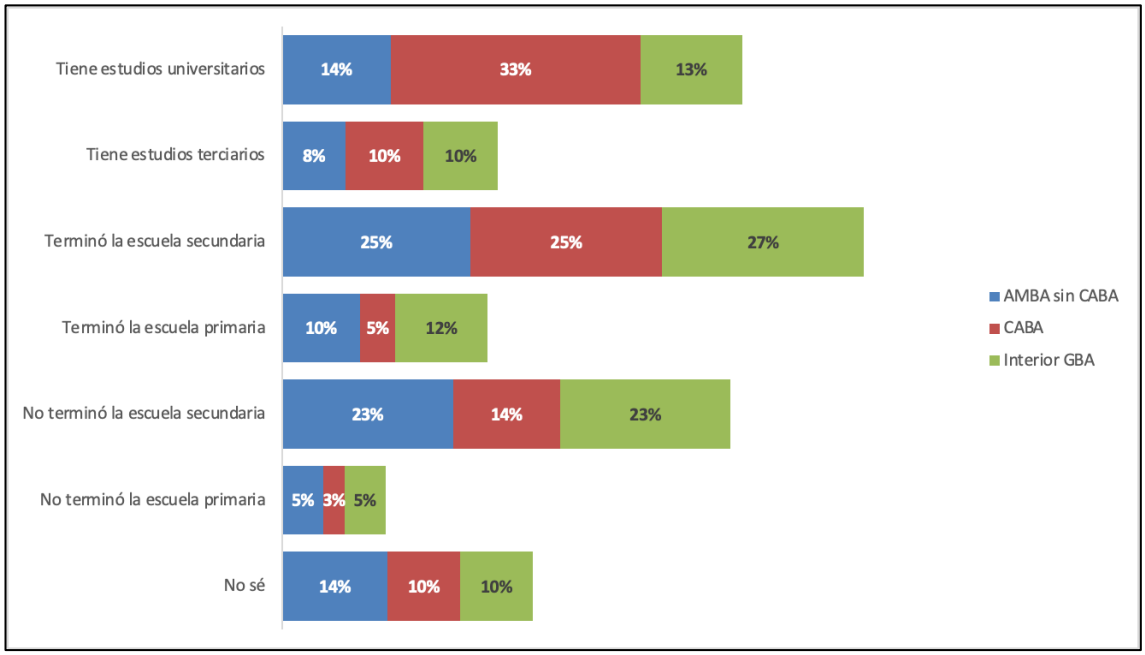


Gráfico 2: Nivel Educativo del Padre (Ap8)



4.2 Exploración

En la base seleccionada se observa un 44% de los alumnos del sexo masculino y el resto femenino. El 97% de las escuelas encuestadas pertenecen al ámbito Urbano, siendo un 53% del sector Privado. Pero al separar el Interior de GBA con las localidades de AMBA, se observa que el interior cuenta con una mayor cantidad de escuelas públicas, ya que solo el 33% pertenece al ámbito privado. El censo Aprender establece como escuela Rural a toda población que tenga menos de dos mil habitantes (Aprender, 2016). Con respecto a la edad de los estudiantes, el 88% tiene entre diecisiete y dieciocho años, lo que tiene sentido ya que estamos trabajando sobre el último año de secundaria, y el restante 12% se encuentra entre los diecinueve y veinte años de edad.

En términos económicos, como se puede ver en el gráfico a continuación, más del 40% se encuentran en la clase Media. Y al separar la Provincia de Buenos Aires de Capital Federal (Tabla 1), se observa una mayor proporción de Clase Alta en la Capital, mientras que la Provincia de Buenos Aires tiene un 69% de su muestra en el percentil del medio. Al hacer un acercamiento de este indicador, por subcategoría, se observan diferencias entre CABA y GBA, pero no entre Interior de GBA y GBA.

Gráfico 3: Nivel Socioeconómico Buenos Aires (isocioa)

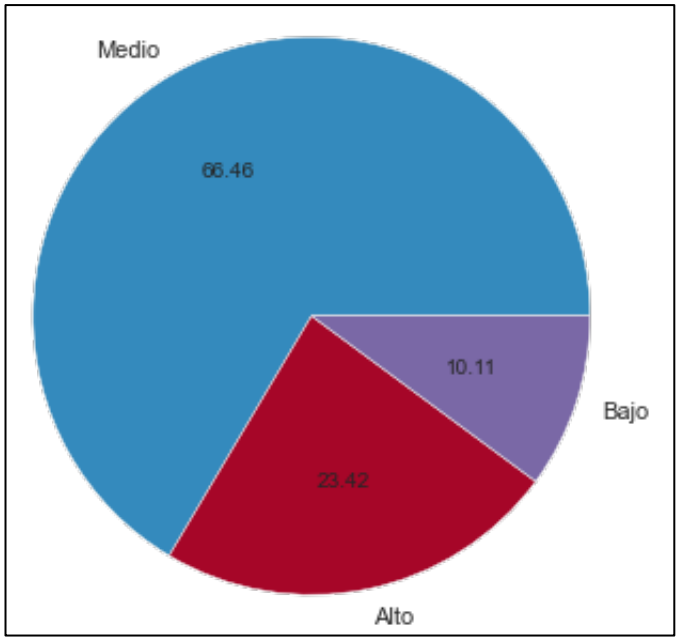


Tabla 1: Nivel Socioeconómico por Subcategoría (isocioa)

Sub Región	**Nivel Socio económico**	**%**
CABA	Bajo	5,2%
	Medio	54,5%
	Alto	40,3%
GBA	Bajo	11,6%
	Medio	68,6%
	Alto	19,8%
Interior GBA	Bajo	9,9%
	Medio	70,7%
	Alto	19,4%

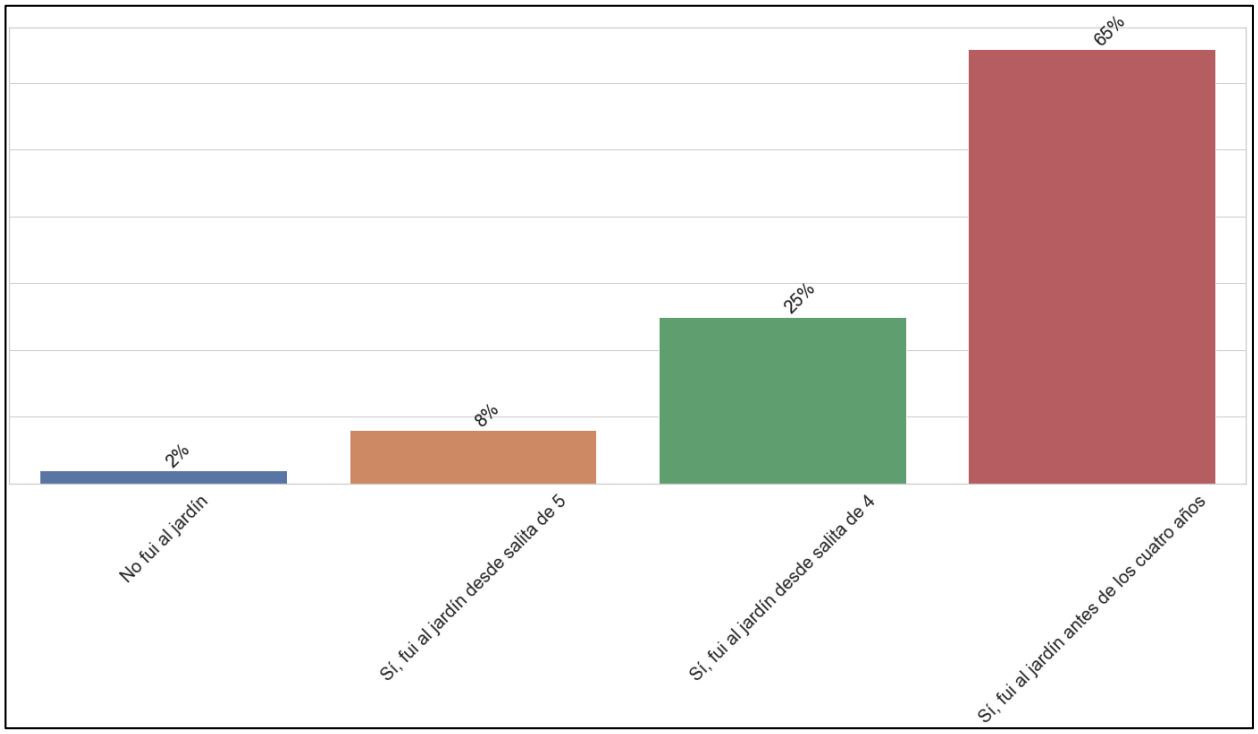
Es importante destacar, que un 26% del *dataset* afirmó que en su hogar cuentan con la Asignación Universal por Hijo o algún tipo de Programa Social, pero como se muestra a continuación, a medida que nos alejamos de CABA el porcentaje se incrementa.

Tabla 2: ¿Alguien de tu familia recibe Asignación Universal por Hijo u otro Programa Social? (Ap6)

Sub región	**Respuesta**	**%**
CABA	Si	13,6%
	No	86,4%
GBA	Si	21,1%
	No	78,9%
Interior GBA	Si	34,9%
	No	65,0%

Hasta el 2014, en Argentina el Jardín de Infantes era obligatorio a partir de los cinco años, pero la Ley Argentina estableció la obligación de comenzar el Jardín a partir de los cuatro años de edad (Senado y Cámara de Diputados Nacional, 2014). Teniendo en cuenta esto, un 90% de los encuestados afirmó haber ido al Jardín desde los cuatro años o antes, en el siguiente gráfico se puede ver la distribución de la edad en la que se comenzó el jardín.

Gráfico 4: ¿A qué edad comenzaste el jardín? (Ap14)



Por otro lado, en cuanto al grado de repitencia, un 5% afirmó haber repetido en Primaria y un 17% en Secundaria. Y como se puede ver a continuación, no hay grandes diferencias al separar CABA de GBA.

Tabla 3: ¿Alguna vez repetiste en secundaria? (Ap16)

Respuesta	Buenos Aires	CABA
Si	16%	18%
No	84%	82%

Con respecto al ausentismo, en el gráfico cinco se puede relevar que un 40% faltó a clases entre 8 y 17 veces, al verlo por región en el gráfico seis la tendencia se mantiene, pero en AMBA sin CABA se observa un porcentaje mayor en los alumnos que faltaron más de 24 veces. Por otro lado, los alumnos en su mayoría no muestran asistencia a clases de apoyo fuera de la escuela, solo el 17% afirmó haber ido alguna vez.

Gráfico 5: ¿Cuántas veces faltaste a lo largo del año? (Ap19)

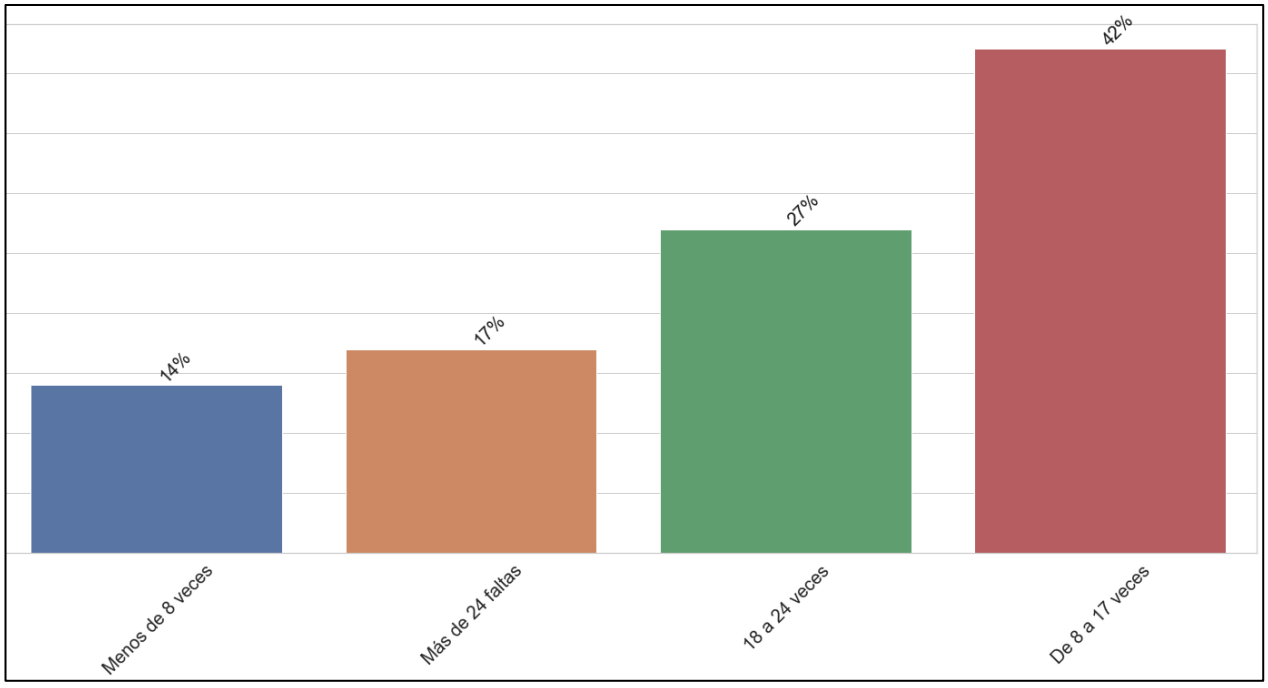
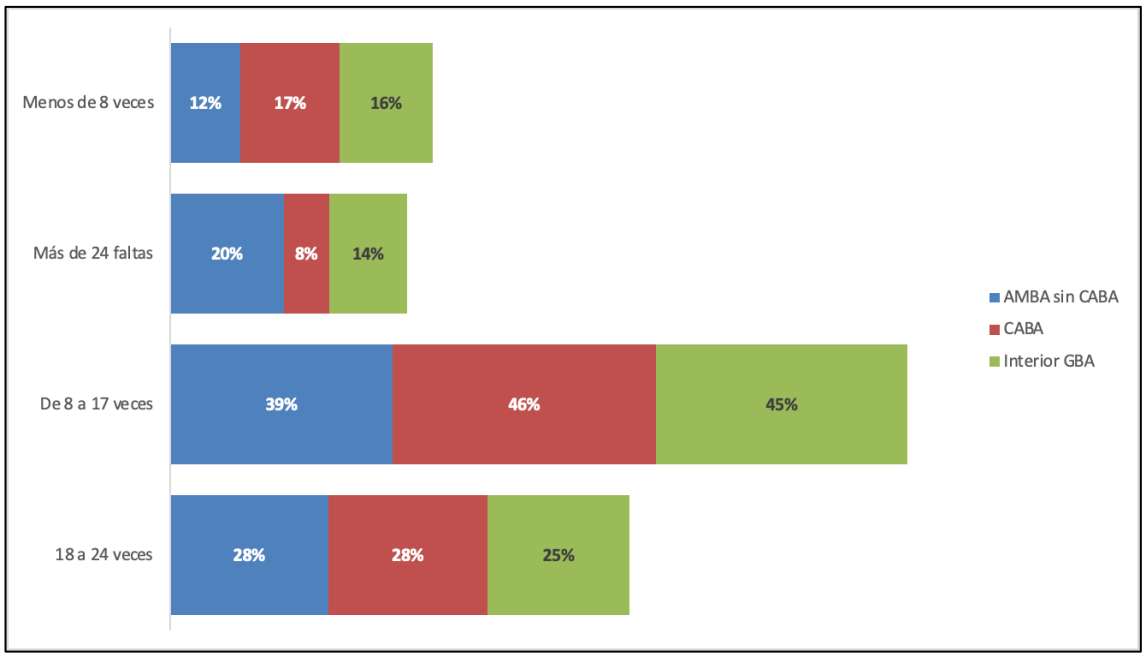


Gráfico 6: ¿Cuántas veces faltaste a lo largo del año? Apertura por Subcategoría (Ap19)



Es importante destacar que el desempeño está comprendido en cuatro niveles: Por debajo del básico, Básico, Satisfactorio y Avanzado. Esta segmentación se construye en base al puntaje del alumno en cada materia, utilizando el método de *Bookmart* ya que es una de las medidas más utilizadas en las pruebas de desempeño escolar. Este método establece los valores correspondientes al límite inferior y máximo de cada nivel, cabe destacar que en Aprender se hizo una modificación al modelo original ya que se sumó un nivel más que es “Por debajo de Básico” esto surge por la necesidad de hacer foco en segmento que no cumplía con los niveles satisfactorios, este grupo comprende a aquellos alumnos que tienen un 25% o más de distancia al límite inferior del nivel “Satisfactorio” (Aprender, 2016).

Al analizar el desempeño de las materias Lengua y Matemática, tal como se observa en los siguientes dos gráficos, hay una notable diferencia entre las áreas. Con respecto a Lengua, un 60% de los alumnos se encuentra en los niveles Avanzado o Satisfactorio y un 19% por debajo del Nivel Básico. En cambio, en Matemática un 39% se encuentra en los niveles Avanzado o Satisfactorio y un 32% tiene un nivel por debajo del Básico. Si observamos la tabla cuatro se pueden ver las sub regiones con peor nivel de desempeño en Matemática que son GBA y el Interior de la Provincia. Mientras que, CABA tiene el porcentaje más alto de nivel Avanzado con un 15% de alumnos. Cabe destacar que el desempeño de cada área tiene un nivel de correlación de Spearman entre 20% y 35%.

Gráfico 7: Nivel de Desempeño en Lengua (ldesemp)

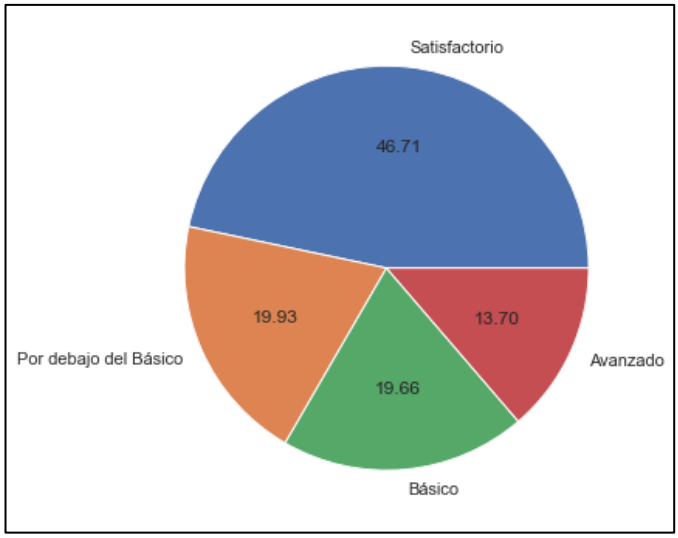


Gráfico 8: Nivel de Desempeño en Matemática (mdsemp)

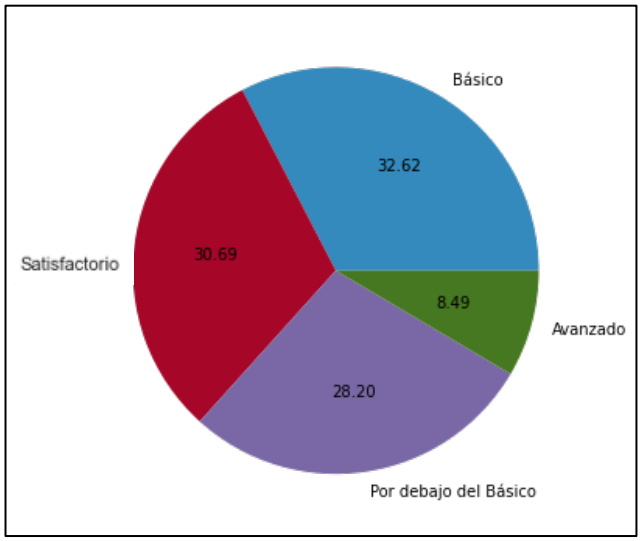


Tabla 4: Nivel de Desempeño en Matemática por Sub Región

Nivel	**CABA**	**GBA**	**Interior GBA**
Debajo del Básico	18,2%	36,7%	33,0%
Básico	24,3%	29,1%	28,7%
Satisfactorio	41,5%	27,4%	30,8%
Avanzado	16,0%	6,8%	7,5%

Capítulo 5: Objetivos Generales

5.1 Percepción de la Tecnología

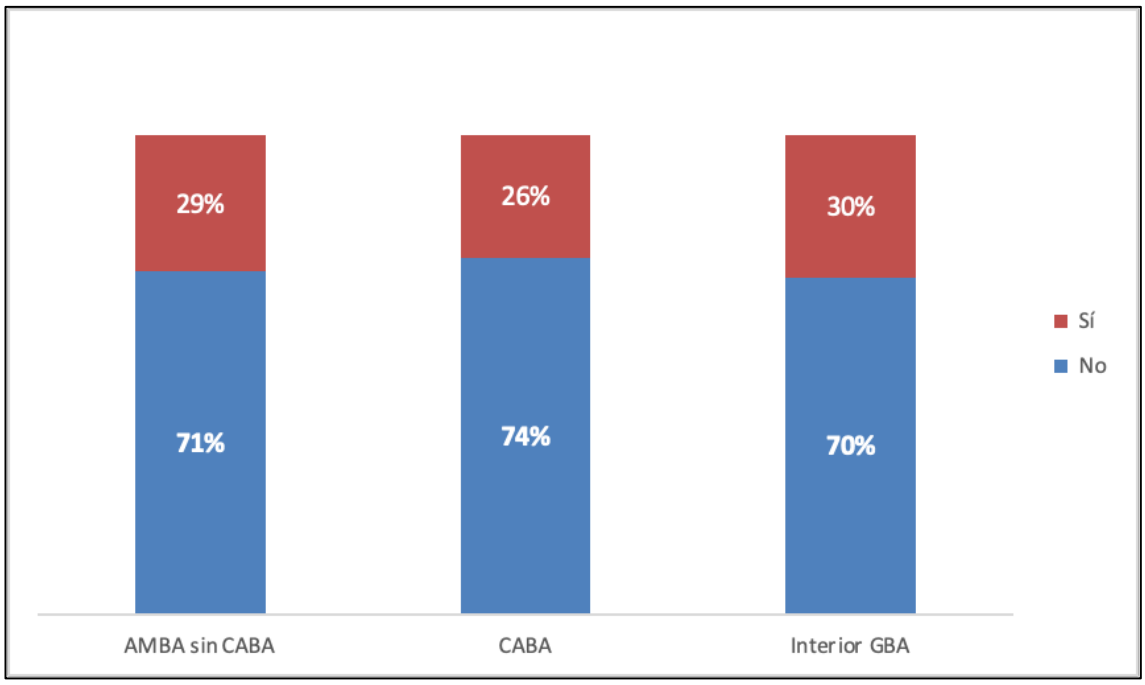
En primer lugar, al observar el acceso a internet, un 84% afirma tener conexión a internet en su hogar, siendo Cañuelas, General Guido, General Rodriguez, Jose C. Paz y Villarino los cinco municipios con menor acceso a la red, donde al menos el 33% negó tener conexión. Estos datos se encuentran por encima del promedio País, donde el 64% de los hogares argentinos tiene internet Fijo (Ente Nacional de Comunicaciones, 2020). Luego al relevar la cantidad de estudiantes que tienen una computadora de escritorio o laptop, un 70% es positivo, pero como se puede ver en la siguiente tabla hay diferencias entre CABA Y GBA. Y prácticamente todos los encuestados afirman tener celular propio, solo un 4% respondió negativamente que se encuentra en GBA.

Tabla 5: ¿En tu casa tenes Notebook o PC de Escritorio? (Ap40a y Ap40b)

Región	**Variable**	**Respuesta**	**%**
CABA	Notebook	No	23%
		Si	77%
	PC Escritorio	No	33%
		Si	67%
GBA + Interior	Notebook	No	32%
		Si	68%
	PC Escritorio	No	33%
		Si	67%

Al relevar el uso que los estudiantes le dan a la tecnología, hay un importante número de alumnos que utiliza la tecnología para actividades escolares, tales como buscar información para la escuela, estudiar o leer artículos. Como se puede ver a continuación, el 70% de los niños afirma que en la Escuela no permiten el uso de la tecnología, aunque un 55% de los mismos afirman que las clases resultan más entretenidas y qué perciben un mayor aprendizaje si se utiliza la computadora.

Gráfico 9: ¿Te permiten usar el celular en el aula? Apertura por sub región (Ap45)



Al unir las variables desempeño en Matemática y Lengua con el uso del celular en la escuela, solo un 25% de los alumnos con desempeño avanzado afirman utilizar el celular en la escuela, ambas materias reflejan probabilidades similares. Pero, si observamos a los alumnos que afirman utilizar la computadora al menos una vez al año dentro del aula, nos encontramos con una mejora de desempeño en Matemática. Al tomar los alumnos que respondieron que nunca han usado la computadora en el aula, un 58% tiene nivel básico o por debajo del básico, siendo un 30% por debajo del básico. Por otro lado, al seleccionar aquellos que respondieron afirmativamente, el 51% está en un nivel básico o menos, siendo el 26% por debajo del básico. Esto puede dar un indicio de la relación entre la tecnología y el desempeño académico que más adelante fue observada.

Tabla 6: ¿Usas tu computadora o celular para…? (Ap43)

Variable	Missing	Sí	No
Buscar información para la escuela	0,00%	71,55%	28,45%
Compartir fotos o videos en internet	0,00%	68,21%	31,79%
Comunicarte con familiares	0,00%	75,57%	24,43%
Comunicarte con personas que no conozcas	0,00%	16,24%	83,76%
Estudiar temas para la escuela	0,00%	49,43%	50,57%
Jugar	0,00%	55,76%	44,24%
Leer artículos o libros digitales	0,00%	41,73%	58,27%
Mandar mensajes a tus amigos	0,00%	88,81%	11,19%
Navegar por redes sociales (Facebook, Google+, Twitter u otras)	0,00%	80,65%	19,35%
Sacar fotos	0,00%	77,63%	22,37%
Seguir a personas conocidas	0,00%	58,39%	41,61%
Ver videos	0,00%	84,16%	15,84%

En base a los valores de la tabla anterior, se puede entender cómo usan la tecnología los alumnos fuera de la escuela. Al explorar esta variable los valores “No” eran del 0%, concentrándose todos entre “Sí” y “*Missing*”. Según lo analizado en variables anteriores, ninguna presentó este patrón, por lo tanto se asume que las respuestas “*Missing*” son negativas.

Cabe destacar, que aquellos que utilizan la tecnología para estudiar temas de la escuela, tienen una tendencia creciente en el desempeño de Matemática y Lengua, ya que un 65% de los alumnos Avanzados de Lengua afirman utilizarlo para este fin y por su lado Matemática concentra un 60%.

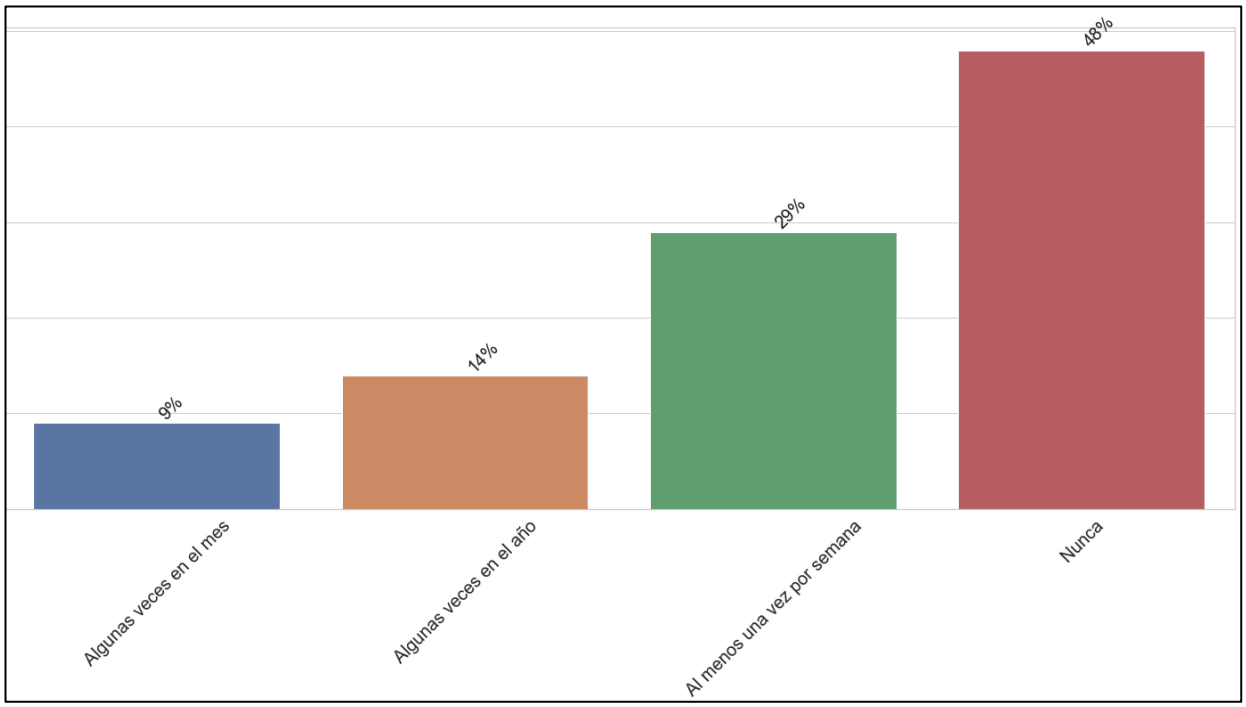
Ahora bien, como se puede ver en la siguiente tabla que describe el uso de la tecnología en los tiempos libres, se presenta una gran parte de la muestra que afirma utilizarla para buscar información, resolver tareas y trabajar proactivamente con sus compañeros

Tabla 7: En tu tiempo libre, ¿con qué frecuencia usas la computadora o el celular para… (Ap44)

Variable	**Al menos una vez por semana**	**Algunas veces en el mes**	**Algunas veces en el año**	**Nunca o casi nunca**
Buscar información en internet.	55%	31%	9%	4%
Escribir trabajos prácticos o resolver tareas	34%	42%	15%	10%
Conectarte con tus compañeros y resolver tareas	43%	29%	15%	13%
Hacer videos o sacar fotos para trabajos de la escuela	24%	27%	26%	23%

Por otro lado, si observamos el siguiente gráfico, es llamativo que hasta en las materias informáticas los alumnos perciben un poco uso de la tecnología durante sus clases. Y por supuesto, el porcentaje de uso disminuye cuando se trata de materias no relacionadas a la tecnología, siendo un 12% menor al observado en Informática.

Gráfico 10: ¿Con qué frecuencia usas la computadora para trabajar en clase de Informática? (Ap47b)



Al hacer foco en los momentos donde se usa la tecnología en las clases, tal como presenta la tabla a continuación, se pueden observar que las actividades más usuales son buscar información en internet, producir textos y documentos, resolver cálculos y trabajar en colaboración con compañeros. Mientras que las menos utilizadas son escribir un programa informático en lenguajes específicos y simulaciones.

Tabla 8: ¿Qué tipo de actividades te proponen hacer con la computadora cuando estás en horario de clases? (Ap48)

Variable	**Al menos una vez por semana**	**Algunas veces en el mes**	**Algunas veces en el año**	**Nunca o casi nunca**	**No corresponde**
Buscar y seleccionar información en internet	22%	20%	15%	15%	28%
Leer en la pantalla un texto escrito por el docente	11%	15%	14%	30%	29%
Producir textos y documentos	18%	20%	18%	15%	29%
Producir recursos multimedia (sacar fotos, editar imágenes o videos)	10%	14%	19%	28%	29%
Responder cuestionarios en la computadora	7%	10%	16%	37%	29%
Jugar con videojuegos educativos	3%	4%	8%	56%	30%
Trabajar en colaboración con tus compañeros	17%	18%	18%	17%	29%
Chatear, usar redes sociales (Facebook, Twitter) o blogs	7%	4%	5%	55%	29%
Realizar cálculos y resolver problemas	11%	12%	17%	31%	29%
Usar simulaciones	5%	6%	9%	50%	30%
Escribir un programa informático mediante el uso de lenguaje especializado	8%	7%	11%	45%	29%

Es importante destacar que estas variables mantienen un alto nivel de correlación de Spearman con las preguntas que relevan el sentimiento de los alumnos en el aula cuando se usa la tecnología. Si tomamos las respuestas en donde se afirma que al usar la tecnología se aprende más, las principales tareas que se destacan son:

Responder cuestionarios en la computadora, Jugar con video juegos educativos, Realizar cálculos y resolver problemas, Usar simulaciones y Escribir un programa informático mediante un lenguaje especializado. A su vez, gracias a la correlación de Kendall, se puede observar un alto nivel de relación entre el uso de simulaciones y la pregunta sobre si se aprende más al utilizar la tecnología (57,90% de correlación con un p valor de 0).

5.2 Relación entre el Nivel Socioeconómico y el Desempeño Académico

Tal como se estableció desde el inicio, uno de los objetivos de la investigación es relevar la relación entre el Nivel Socioeconómico del alumno y el Acceso a la Tecnología. Por lo tanto, se hará un análisis de Regresión para estudiarlo.

En primer lugar, se definirán como variables *target* el puntaje de Matemática (mpuntaje) y Lengua (lpuntaje). Una vez seleccionado, se generaron los siguientes dos gráficos de las variables que permitió controlar la distribución de las mismas y así validar que sea normal.

Gráfico 11: Distribución Puntaje de Evaluación Lengua

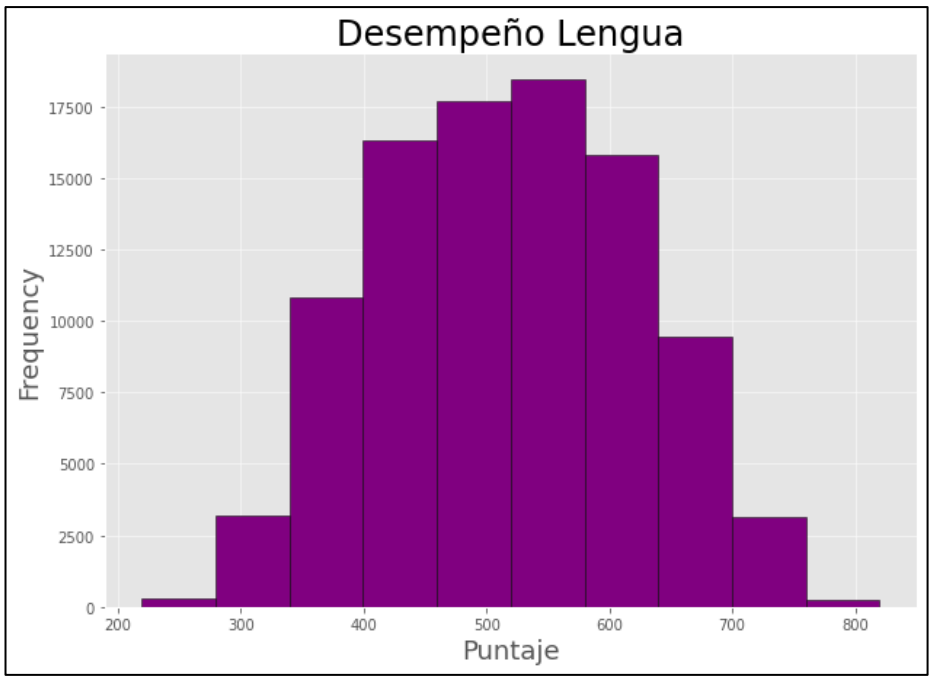
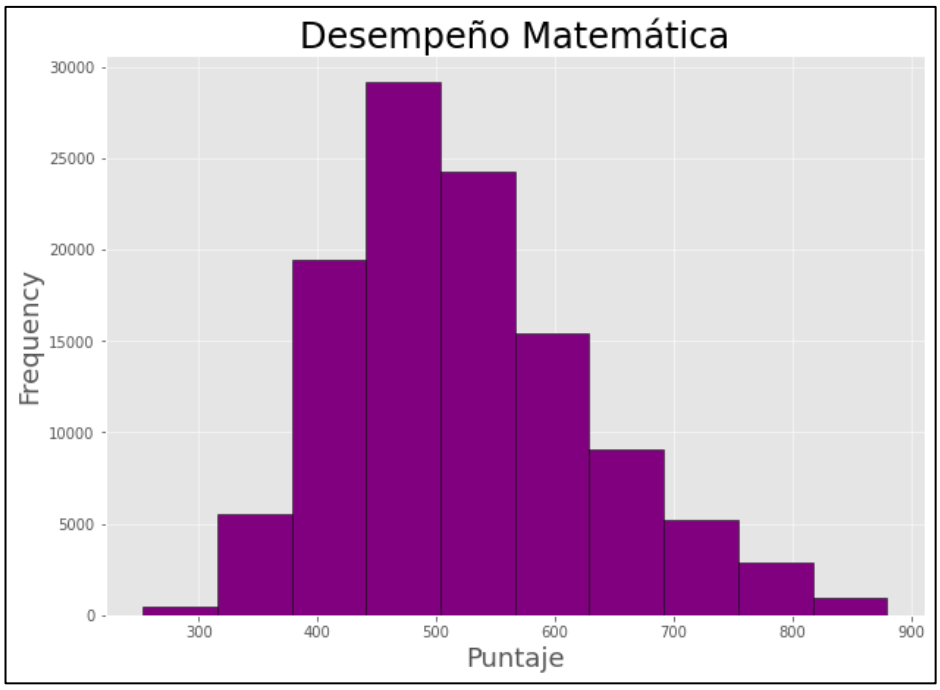


Gráfico 12: Distribución Puntaje de Evaluación Matemática



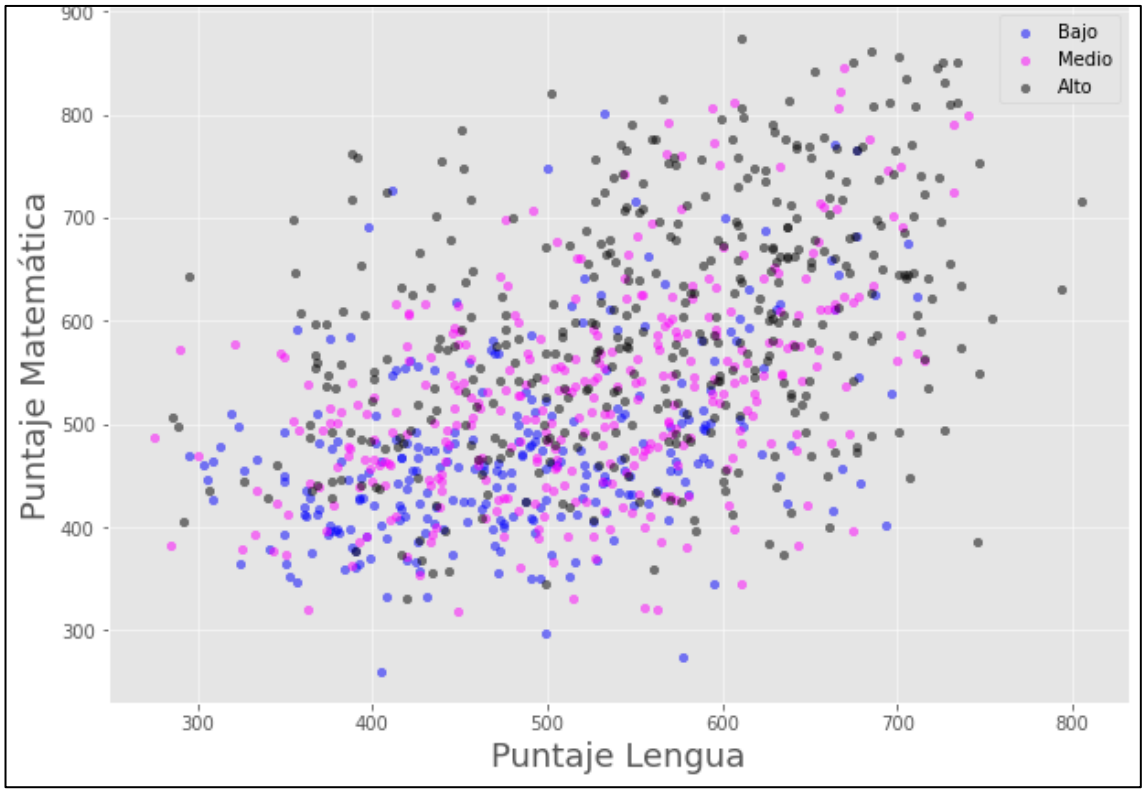
Una vez observada la distribución de las variables *target*, se analizó la proporción de alumnos por desempeño, según el nivel socio económico y el promedio del puntaje para cada materia. Dado que, al avanzar de cuartil económico, la distribución de los desempeños se concentra en los niveles satisfactorios, a priori se puede afirmar una relación empírica entre el status social y el desempeño. A su vez, el puntaje promedio es creciente a medida que se avanza en la escala social. Tal como se observa a continuación esto sucede tanto en Lengua como en Matemática.

Tabla 9: Nivel de Desempeño en Lengua y Matemática, según el Nivel Socioeconómico.

Materia	**Desempeño**	**Nivel Socioeconómico**		
		Bajo	**Medio**	**Alto**
Lengua	Por debajo del Nivel Básico	30%	20%	13%
	Básico	29%	21%	13%
	Satisfactorio	37%	48%	50%
	Avanzado	4%	12%	24%
	Puntaje Promedio	470,05	512,43	553,69
Matemática	Por debajo del Nivel Básico	54%	34%	16%
	Básico	30%	30%	21%
	Satisfactorio	15%	30%	42%
	Avanzado	1%	6%	21%
	Puntaje Promedio	466,22	512,42	585,13

Como se puede observar en el siguiente gráfico, al tomar una muestra aleatoria de 550 registros por *status* social, se obtienen datos con un alto nivel de dispersión

Gráfico 13: Relación entre el Nivel Socioeconómico y el puntaje de Matemática y Lengua.



Ahora bien, al correlacionar las variables con el método de Pearson, la relación del Nivel Socioeconómico y Lengua es de 22%, con un p-value de 0% y la relación con Matemática es de 33%, con el mismo valor de p. Y al separarlo por Subcategoría, se mantiene la misma relación, siendo 3 puntos más fuerte la relación en AMBA sin CABA.

Es importante mencionar que el desempeño de las materias se encuentra relacionado, y al relevar la dispersión de los datos por nivel socio económico se encuentra un coeficiente de variación promedio de 18%.

Capítulo 6: ¿Qué factores influyen en el Desempeño Académico?

6.1 Variable Target

En base a lo observado en los capítulos anteriores, ya se reconocieron los datos, entendiendo su procedencia y veracidad. Luego se agregaron datos relevantes para trabajar sobre un análisis descriptivo, que permitieron encontrar patrones relevantes para describir los datos (Gregory, Padhraic, Ramasamy & Usama, 1996).

Ahora, se busca trabajar sobre el objetivo principal, para entender qué factores explican el desempeño de los alumnos de secundaria en Buenos Aires, dentro de las áreas de Matemática y Lengua.

Cabe destacar que la variable *target* puede ser el puntaje obtenido en los exámenes o el desempeño, dado que la segunda se construye en base a la primera, se va a seleccionar una u otra dependiendo el modelo predictivo de mejor rendimiento. Con la intención de obtener un mejor resultado, se corrió el mismo modelo, pero trabajando las variables *target* por separado, de esta manera se mitigó el sesgo del análisis ya que Lengua y Matemática tienen una correlación de 34%.

Tal como se observó en los Gráficos 11 y 12, las dos variables *target* se distribuyen normalmente. Esta distribución se mantiene al extrapolarlo en el Nivel Socioeconómico.

6.2 Feature Importance

Luego de medir diferentes correlaciones y debido a que las variables predictoras son ordinales numéricas, el mejor enfoque de correlación es Kendall ya que el nivel de relación depende en las coincidencias entre las variables.

$$\tau = \frac{(\text{número de pares concordantes}) - (\text{número de pares discordantes})}{\binom{n}{2}}.$$

Dado que no se observan variables fuertemente correlacionadas, el máximo obtenido es de 30% (excluyendo el desempeño entre las áreas), se procederá a evaluar otros modelos que permitan entender qué variables predicen mejor al desempeño de los alumnos. Otro punto importante es que la relación entre los desempeños de otras materias no está fuertemente relacionado. En las siguientes dos tablas se presentan las

variables del cuestionario anexo que tienen más de 18% de correlación al desempeño definido como *target*.

Tabla 10: Principales correlaciones en Lengua (mayor o igual a +-18%)

Variable	**lpuntaje**	**Variable**	**ldesemp**
ldesemp	82,7%	lpuntaje	82,7%
cndesemp	46,4%	cndesemp	49,5%
cnpuntaje	45,1%	cnpuntaje	47,2%
csdesemp	43,7%	csdesemp	46,3%
cspuntaje	42,3%	cspuntaje	44,1%
mdesemp	34,6%	mdesemp	35,9%
mpuntaje	32,5%	mpuntaje	33,6%
sector	21,5%	sector	22,8%
Tenes Smartphone	20,7%	Tenes Smartphone	21,5%
Usas computadora para leer artículos digitales	18,6%	Usas computadora para leer artículos digitales	19,5%
		Tenes Materias Previas del año pasado	18,7%

Tabla 11: Principales correlaciones en Matemática (mayor o igual a +-18%)

Variable	**mpuntaje**	**Variable**	**mdesemp**
mdesemp	84,4%	mpuntaje	84,4%
cnpuntaje	37,1%	cnpuntaje	39,6%
cndesemp	36,3%	cndesemp	38,9%
csdesemp	34,6%	csdesemp	37,1%
cspuntaje	34,5%	cspuntaje	36,9%
ldesemp	33,6%	ldesemp	35,9%
lpuntaje	32,5%	lpuntaje	34,6%
Tenes Smartphone	26,6%	Tenes Smartphone	28,8%
isocioa	26,1%	isocioa	28,2%
sector	25,8%	sector	28,0%
Tenes Materias Previas del año pasado	19,5%	Tenes Materias Previas del año pasado	21,2%
Nivel de Estudios de la Madre	18,7%	isocioa	28,2%
En la clase de Matemática entiendo rápido	-20,7%	Nivel de Estudios de la Madre	20,2%
Te fue difícil resolver la prueba de Ciencias Sociales	-28,0%	En la clase de Matemática me va bien	-19,2%
		En la clase de Matemática entiendo rápido	-22,4%
		Te fue difícil resolver la prueba de Matemática en Aprender	-30,2%

Continuando con el análisis de *feature importance*, se comenzó trabajando con un algoritmo de *Random Forest*, que corresponde al grupo de modelos aplicados para clasificación y es también utilizado para medir la relevancia de las variables de un *dataset*. A su vez, no es muy sensible a *outliers* y no requiere distribución normal en todas las *features*. Esta primera aproximación, nos permitirá entender qué sucede cuando agregamos variables al modelo, si alguna de ellas contribuye a la predicción o no (Breiman, 2001). Para esta aplicación se utilizó la librería de Sklearn, parametrizando un 25% del *dataset* en prueba y el resto para entrenamiento. El resultado obtenido fue con un *accuracy* de 47%, tanto en Lengua como Matemática. Ahora bien, si se analiza la importancia de cada variable, las principales predictoras, que no alcanzan el parámetro de 0.012 del modelo, son el Nivel Educativo del Padre y Madre, la Cantidad de Personas que Viven en el Hogar y la Cantidad de Habitaciones. Lo llamativo es que el conjunto de estas variables construye el Nivel Socioeconómico y éste tiene una relevancia inferior al momento de predecir el desempeño en Lengua. Por otro lado, si se aplica *Permutation Importance*, donde la importancia se mide a través del incremento en el error predicho al modificar las variables predictoras (Molnar, 2020), toma relevancia el sector y si en la casa tiene Smartphone o no.

Con respecto a Matemática, el *accuracy* alcanzado en el testeo es de 46%, y las variables con mayor puntaje de predicción coinciden con Lengua. Como primera aproximación, se podría afirmar que un solo modelo podría inferir en ambas áreas, pero la correlación entre ambas es de 32%.

En una segunda etapa, se aplicó un algoritmo de *XGBoost*, ya que es considerado un modelo robusto y con alta escalabilidad en diferentes escenarios (Chen & Gestri, 2016). Esta metodología consiste en ir agregando *features* que mejoren la *performance*.

Para este caso, se definieron dos indicadores de importancia, en primer lugar, el peso de la variable, que es la cantidad de ocurrencias para cada árbol. Al aplicarlo para predecir el desempeño en Lengua, donde al excluir el Municipio y el auto concepto de la evaluación Aprender, vuelve a figurar la variable Ap7 que hace referencia al Nivel Educativo de la Madre. En segundo lugar, el indicador observado es la ganancia, que mide la ganancia promedio que aporta la variable en cada *split*. Aquí sí, aparecen variables nuevas, tales como el Sector, que puede ser público o privado y luego la variable Ap15, que releva la cantidad de veces que se repitió primaria. Y muy cerca de ésta, aparece la variable Ap40e, donde se pregunta si el alumno tiene Smartphone en la casa o no. Finalmente, el *accuracy* alcanzado fue de 48%. En los próximos cuatro gráficos se reflejan las variables más relevantes por materia y según la metodología definida.

Gráfico 14: Peso de las Variables (Lengua)

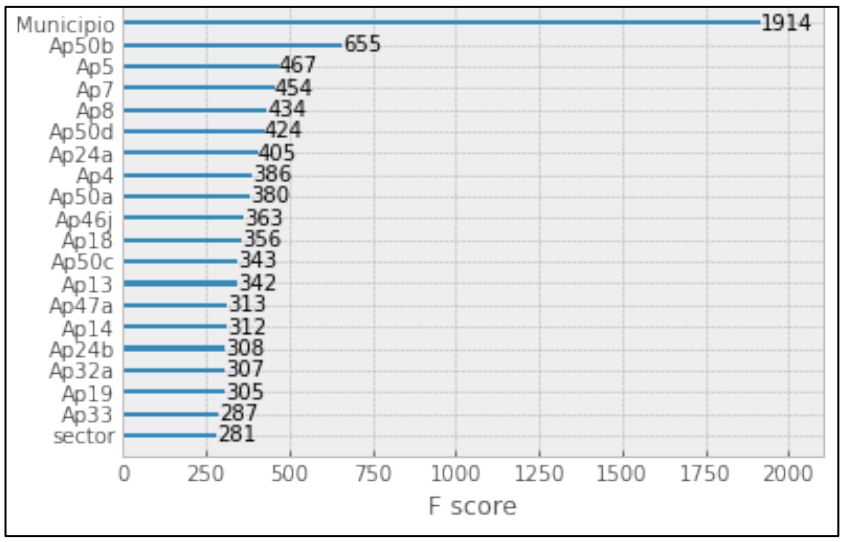
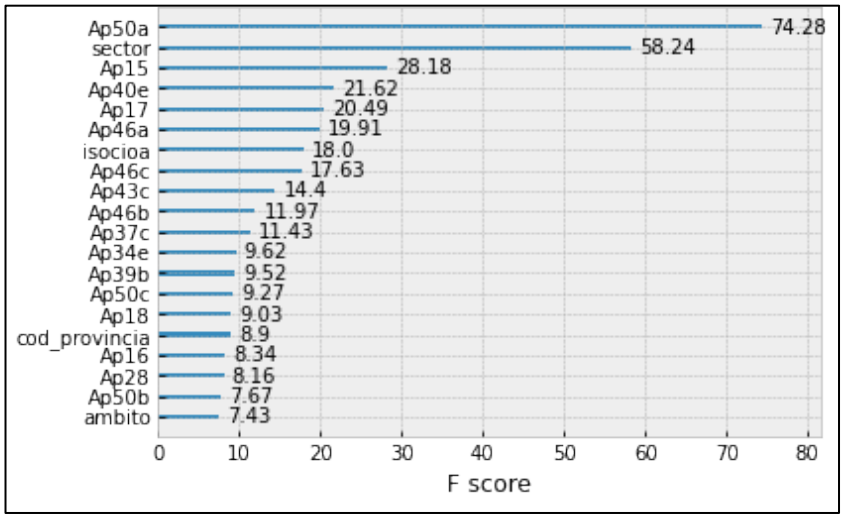


Gráfico 15: Ganancia de las Variables (Lengua)



Por otro lado, en Matemática, si se releva el peso de las variables, el resultado es similar a Lengua, pero ahora también el Nivel Educativo del Padre toma mayor relevancia. Pero en el caso de la ganancia, sí aparecen nuevas variables, tales como el Nivel Socio Económico del alumno y si tienen materias previas del año pasado o no. También se mantiene la variable Ap15, donde se pregunta si el alumno repitió en primaria. Por último, en esta materia el *accuracy* resultante también fue de 48%.

Gráfico 16: Peso de las Variables (Matemática)

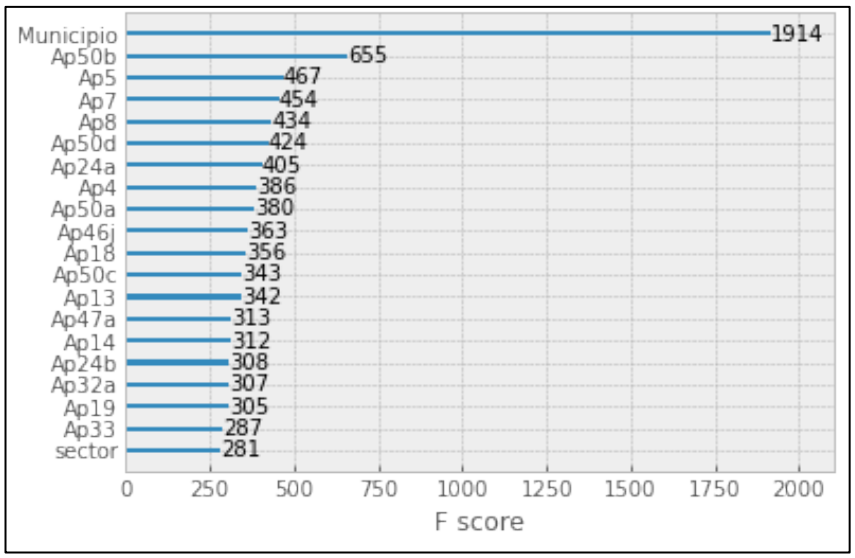
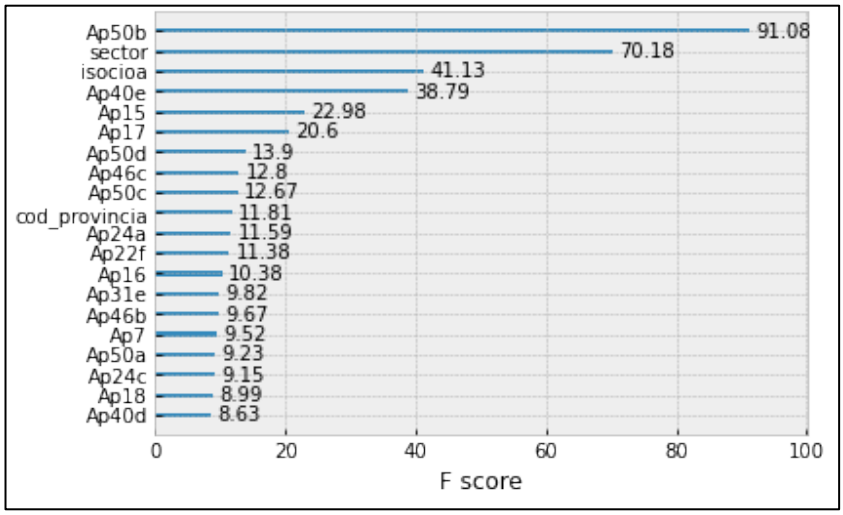


Gráfico 17: Ganancia de las Variables (Matemática)



6.3 Modelo Seleccionado

Retomando con el objetivo principal de la investigación, donde se busca predecir el nivel de desempeño de los alumnos en Matemática y Lengua, cuyos valores posibles son los siguientes:

Tabla 12: Resultados Posibles en Desempeño Académico

1	Por debajo del Nivel Básico
2	Básico
3	Satisfactorio
4	Avanzado

Para lograr esto, el valor que se necesita predecir es el puntaje que obtiene cada alumno en el examen de Aprender para cada área.

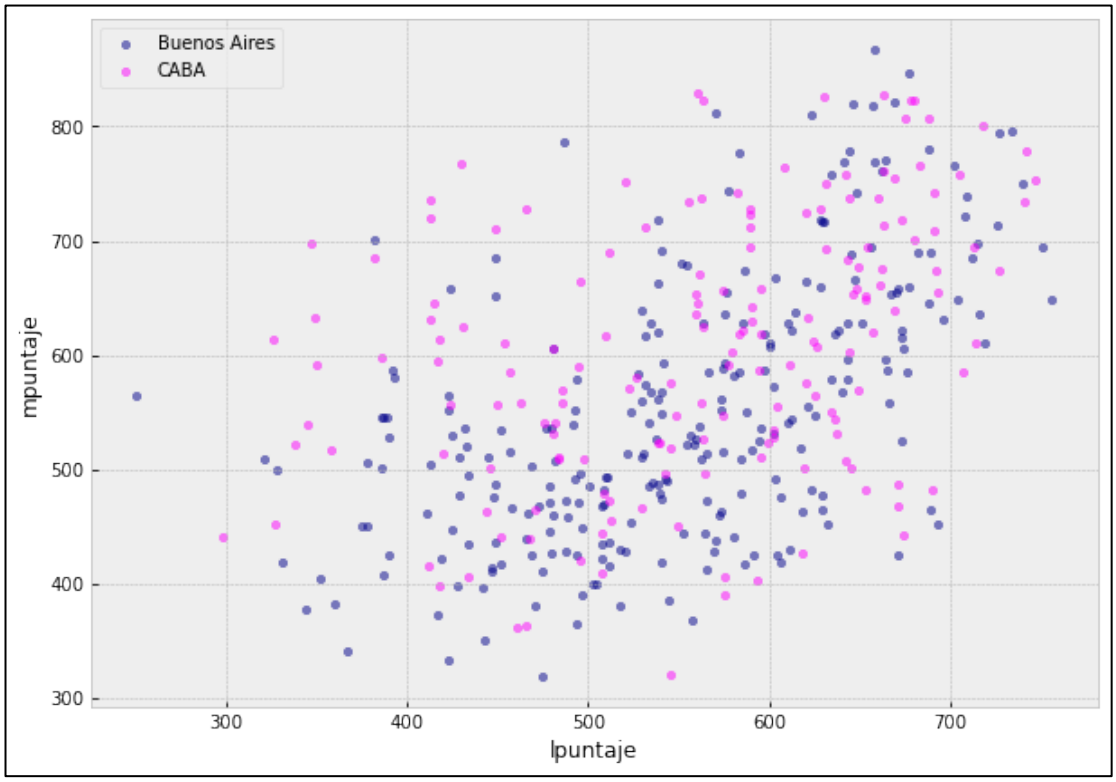
Al tratarse de un trabajo de investigación que utilizó datos de Ciencias Sociales pueden aparecer variables *non-stationary* que son *features* cuyo resultado de análisis depende directamente de dónde se midió. Por este motivo es importante analizar el nivel de error captado por variables con frecuencias diferentes entre regiones geográficas. El *dataset* cuenta con la mínima expresión de Municipio y hasta el momento todas las conclusiones que se fueron abordando se trataron de medidas globales, analizando el dato como un todo y cuando se trabajó de forma local fue dentro de la variable agregada *Subcategory* (Brunsdon, Charlton & Fortheringham, 2002).

Antes de seleccionar el modelo se corroboró si los datos requerían de una solución de algoritmos geo referenciados. Al tratarse únicamente de la Provincia de Buenos Aires, la primera apertura para hacer la validación se hizo entre CABA, AMBA sin CABA y el Interior de GBA, siendo CABA el 15% de la base. Para hacer la validación se tomaron las variables que se destacaron en relevancia a lo largo de la investigación, en primer lugar, al observar el nivel socioeconómico la mayoría está comprendido dentro del nivel medio para las tres regiones, pero en CABA se presenta una frecuencia mayor en el nivel económico alto. En segundo lugar, el nivel educativo de los padres, tal como se vio en capítulos anteriores en CABA se observa un mayor profesionalismo que en las otras dos regiones.

Por último, al analizar las variables *target*, en Lengua no se encuentra diferencia entre las regiones, pero en Matemática sí hay una pequeña tendencia a niveles

avanzados cuando se observa CABA. Esto último se midió con las variables que relevan el nivel en el que se encuentra cada alumno, pero si se observa la variable del puntaje para cada materia, cómo se puede ver en el gráfico a continuación, no hay diferencias entre la Capital Federal (CABA) y el resto de la Provincia de Buenos Aires.

Gráfico 18: Distribución de Desempeño por Provincia



En base a esto y debido a que la cantidad de casos por Municipio es muy baja para hacer un modelo por cada uno se decidió aplicar un modelo donde se introduzca el *dataset* completo, sin considerar la ubicación geográfica del alumno. El modelo a aplicar será de Regresión, ya que se trata de una variable continua, cuyos valores pueden ser entre cero y mil. Y como resultado final se buscará determinar las clases del desempeño, por lo tanto, el modelo más apropiado es un Árbol de Decisión, específicamente el modelo de Sklearn, que se basa en una versión optimizada de CART (*Classification and Regression Trees*).

En primer lugar, se creó una nueva tabla para eliminar variables que no se han utilizado para el análisis, principalmente geográficas. Y con el objetivo de utilizar

únicamente las variables anexas, se descartan los desempeños y puntajes en el resto de las áreas.

Se comenzó con la materia de Lengua, donde se definió como variable *target* la variable lpuntaje (Puntaje Obtenido en Lengua). El *dataset* de testeo seleccionado fue del 10%. Con este modelo, se alcanzó un score de -41,98%, al tratarse del r2, cuyo valor cercano a uno nos indicaría una predicción perfecta, en este caso será necesario trabajar sobre las variables para mejorar la performance. Con el objetivo de tener un acercamiento a variables destacadas, se observó la importancia de cada una para el modelo, basándose en el Coeficiente de Gini que mide la impureza de cada *feature* a lo largo del árbol, también definido como la probabilidad de minimizar la mala clasificación del árbol (Bhattacharyya, 2019). Finalmente, como se puede ver en la siguiente tabla, las variables de mayor importancia para Lengua que se obtuvieron son:

Tabla 13: Variables más Importantes para Lengua

Variable	**Descripción**	**Gini Index**
Ap50a	Te resultó difícil resolver el exámen de Lengua?	14,23%
Municipio		4,42%
Sector	Estatal o Privado	4,15%
Ap15	Repitió Primaria?	2,41%
Ap8	Nivel educativo Padre	1,42%

En segundo lugar, se corrió el mismo modelo para Matemática, donde se obtuvo un r2 de -22,59%, pero en este caso se observa un árbol menos profundo, con 103 niveles. Tal como refleja la tabla a continuación, si bien algunas variables se mantienen en el top cinco de importancia, aparece una nueva, que es la dificultad para resolver el examen de Ciencias Sociales, sacándole el lugar al Nivel educativo del Padre, que se encuentra seis niveles más abajo con un índice de Gini igual a 0,90%.

Tabla 14: Variables más Importantes para Matemática

Variable	Descripción	Gini Index
Ap50b	Te resultó difícil resolver el Exámen de Matemática?	22,1%
Sector	Estatal o Privado?	4,5%
Municipio		3,8%
Ap15	Reitió Primaria?	2,8%
Ap50d	Te resultó difícil resolver el Exámen de Ciencias Sociales?	1,8%

Tabla 15: Parámetros Utilizados en los Árboles de Decisión

Parámetro	Valor
ccp_alpha	0
criterion	MSE
max_depth	None
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0
min_impurity_split	None
min_samples_leaf	1
min_samples_split	2
min_weight_fraction_leaf	0
presort	deprecated
random_state	None
splitter	Best

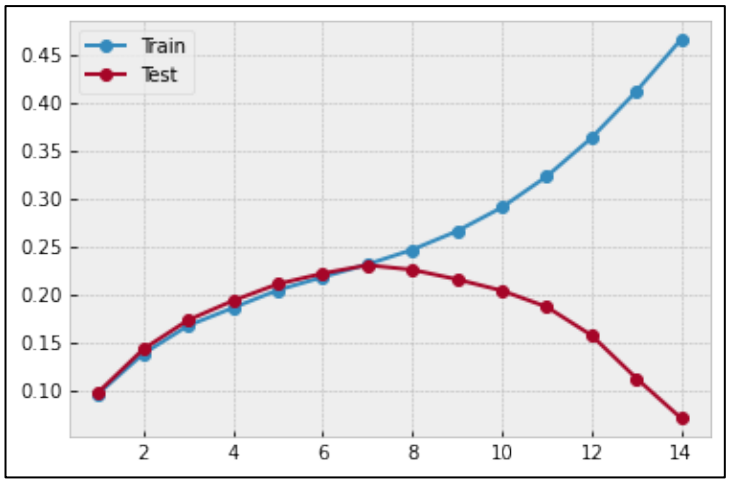
Con los parámetros detallados en la tabla anterior y tal como se ha observado en la *performance* de los modelos, se presenta fuerte tendencia a *overfitting* ya que en entrenamiento alcanza un score casi perfecto, pero cuando se hace el testeo el efecto es negativo. En consecuencia, fue necesario hacer un ajuste en las variables.

6.4 Mejora del Modelo Predictivo

Para lograr un mejor predictor es necesario reducir la cantidad de variables redundantes, para ello se utilizó un modelo de componentes principales. Particularmente se aplicó un análisis factorial ya que permite una interpretación más simple a la hora de entender qué variables contiene cada factor, este modelo arroja como resultado el coeficiente de correlación entre un componente y una variable que representa la parte de la varianza que explica cada factor (Peña, 2002). Específicamente en este caso se trabajó con la metodología "varimax" ya que funciona en diferentes aplicaciones. La cantidad de componentes necesarios para cubrir el 80% de la varianza del *dataset* es de 35.

Una vez que se redujeron las variables se procedió a aplicar el mismo modelo de árboles de decisión con regresión, pero antes se hizo un análisis sobre el r2 de testeo y entrenamiento para ver a partir de qué profundidad el modelo empezaba a hacer *overfitting* y así definirlo en los parámetros. Sin hacer este control, los árboles alcanzaban rápidamente una profundidad de cincuenta.

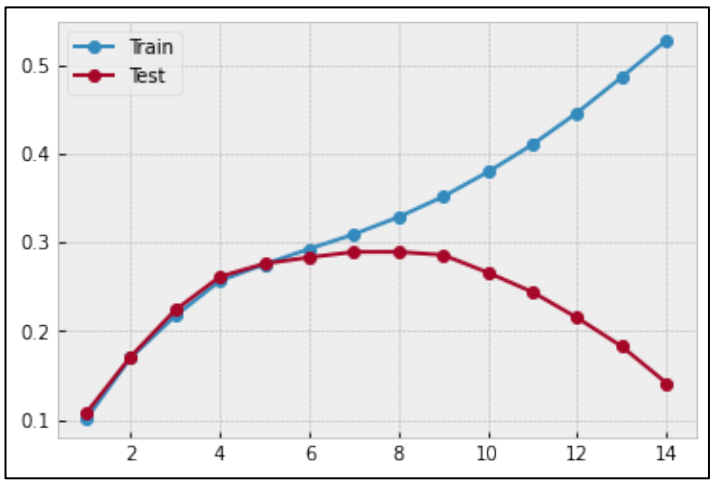
Gráfico 19: Búsqueda de Parámetros para Lengua:



Tal como se puede observar en el gráfico anterior el nivel óptimo para Lengua es con una profundidad de siete. De esta manera se logró alcanzar un r2 de 21,74%, si bien aún no es un modelo con exactitud elevada se ha revertido el valor negativo que

se obtenía antes de aplicar componentes principales. Al entrar en el modelo se puede ver que los factores con mayor peso para la predicción, según el Coeficiente de Gini, es la facilidad para resolver el examen Aprender junto con la trayectoria escolar donde se mide si el alumno repitió alguna vez, si tiene materias previas y si fue al jardín de infantes. También aparece el sector de gestión de la escuela y si tienen Smartphone propio o no. El nivel de estudio de los padres sigue siendo de las variables más relevantes, pero ahora se ubica en una posición seis.

Gráfico 20: Búsqueda de Parámetros para Matemática:



Al observar el gráfico anterior se puede ver que en el caso de Matemática el mejor nivel de profundidad es en nueve ya que obtenemos un r2 de 28,90% en testeo. Aquí aparece nuevamente el factor que releva el sector de gestión de la escuela y si el alumno tiene Smartphone propio, pero en este caso tiene más pesos que para el modelo de Lengua. En menor escala este efecto sucedía también en la aplicación sin componentes principales. Luego aparece un factor que releva el acceso a las TIC en el hogar, donde se pregunta si tienen conexión a internet, computadoras propias y otros elementos tecnológicos. Por último, el componente que mide el auto concepto del alumno en términos de facilidad durante las clases y cómo se sienten en la escuela tiene relevancia para este árbol.

Capítulo 7: Conclusiones

A lo largo de la investigación se recorrió un camino de exploración, que permitió comprender cómo estaban presentados los datos, siempre en la búsqueda de responder las preguntas formuladas en el inicio: ¿Qué necesitan los alumnos del último año de Secundaria para tener un mejor desempeño en Lengua y Matemática? Por supuesto se trata de un largo recorrido, pero entendiendo qué factores contribuyen al desempeño y observando las relaciones entre las variables del cuestionario anexo, principalmente en las referidas a la tecnología y el nivel socio económico se ha logrado llegar a resultados concisos, que nutrirán al ámbito de la Educación de información relevante.

En una primera instancia, se observó que el acceso a internet (móvil y/o fijo) en la Provincia de Buenos Aires es del 84%, valor superior al promedio País. Mientras que, los alumnos afirman utilizar las tecnologías en sus hogares para actividades escolares, tales como, búsqueda de información, lecturas de artículos o trabajo en equipo. En la escuela el uso de la tecnología es de un 25%. Si bien no se ha encontrado relación estadística entre el uso de la tecnología y el desempeño de los alumnos, sí se observa una correlación entre la percepción de aprendizaje por el alumno dentro del colegio y el uso del celular en las clases.

Con respecto al nivel socio económico, aunque se observa una tendencia hacia niveles satisfactorios a medida que se sube de escala social, no se observan correlaciones con el desempeño académico, pero sí ha sido relevante en el modelo predictor de Matemática. Como se ha mencionado en capítulos anteriores, la correlación alcanzada es de 22% en Lengua y 33% en Matemática. Además, el promedio de puntaje en cada escala social, no presenta variaciones importantes, oscilando entre 8 y 12 puntos básicos. Es importante destacar, que se han encontrado relaciones entre el desempeño académico y variables que componen al nivel socio económico, tales como el Nivel Educativo del Padre o la Madre y la Cantidad de Habitaciones que tiene el Hogar. Con lo cual, se podría decir que dentro del índice de nivel socio económico, hay variables que sí contribuyen a que un alumno tenga mejores resultados que otro.

Aunque no alcanzan niveles de correlación alto (más de 60%), las variables que se destacan en términos de relevancia fueron: La percepción de los alumnos sobre el examen Aprender, donde se pregunta si al alumno le resultó fácil o difícil resolverlo, el grado de repitencia (principalmente en primaria), si tienen o no materias previas de años anteriores y si el colegio es público o privado.

En base al modelo aplicado, si bien no se han alcanzado grandes niveles de r2 se obtuvieron factores que contribuyen al desempeño de los alumnos del último año de secundaria en la Provincia de Buenos Aires, para Lengua y Matemática. Dentro de los

factores destacados en Lengua se encuentra la trayectoria escolar del alumno y a qué edad comenzó a ir al jardín, la educación de los padres y qué tan fácil le resultó resolver el examen de Aprender. Por otro lado, en Matemática las variables más relevantes para el modelo predictor fueron: El acceso a las TIC donde se relevó qué elementos tecnológicos tiene el alumno en la casa y la conexión a internet y el auto concepto del alumno en cuanto a las clases y cómo se siente durante la estadía en la escuela. En las dos materias el factor que mide si es un colegio público o privado tiene relevancia para el algoritmo.

Dicho esto, se puede afirmar que la Educación tiene algunos puntos de mejora para asegurar que un alumno de Buenos Aires tenga buen desempeño en Matemática y Lengua. En primer lugar, entendiendo cómo motivar a aquellos adolescentes que tienen padres con niveles educativos más bajos. Esto debería acompañarse con un programa especial para aquellos alumnos que han repetido algún año de la escuela dado que repetir el año, por sí solo, no trae beneficios en el resultado final que sería el desempeño. En segundo lugar, es importante que el aula adapte nuevas tecnologías de forma medida, si bien no se encuentran resultados que sugieran una relación estricta al desempeño, hay indicios que posicionan a la tecnología como una herramienta complementaria para los niños que podría ayudarlos en algunos aspectos. Este último punto se puede ver con el impacto positivo que tiene la tecnología en el aprendizaje percibido por los alumnos cuando la utilizan, principalmente con el uso de videos juegos o simuladores y con la forma de evaluación a través de cuestionarios en la computadora.

Finalmente, con respecto al cuestionario, es importante que haya una continuidad entre cada año, manteniendo el máximo de preguntas posibles permitiendo trabajar con evoluciones. Es posible que el Nivel Socio Económico necesite de más variables para poder agruparlo y así tener un indicador más relevante en modelos estadísticos. Por último, sería enriquecedor formular más preguntas relacionadas a las que han resultado ser más relevantes en el modelo. Por ejemplo, si el nivel educativo de los padres es relevante, entender si reciben ayuda de ellos para hacer tareas o si de alguna forma los padres influyen en el estudio de sus hijos.

Referencias

Adrogué C. y García A. (2015). *Abandono de los Estudios Universitarios: Dimensión, Factores Asociados y Desafíos para la Política Pública*. Revista Fuentes.

Aprender 2016 (2016), *Glosario*. Ministerio de Educación. Link: https://www.argentina.gob.ar/sites/default/files/glosario_aprender2016.pdf

Aprender 2016 (2016), *Medición del Nivel Socioeconómico.* Ministerio de Educación. Link: https://www.argentina.gob.ar/sites/default/files/nivel_socioeconomico.pdf

Aprender 2016 (2016), *Notas Metodológicas*. Ministerio de Educación. Link: https://www.argentina.gob.ar/sites/default/files/notas_metodologicas_0.pdf

Argentinos por la Educación (2011), *Repositorio de Datos Público*. Argentinos por la Educación. Link: https://gitlab.com/AxEeduc/datos

Babin, J., Carr, J., Griffin, M. y Zikmund, W.G. (2012), *Business Research Methods: with Qualtrics Printed Access Card.* Cengage Learning.

Banco Interamericano de Desarrollo (2018), *Algoritmolandia Inteligencia artificial para una integración predictiva e inclusiva de América Latina*. Editorial Planeta.

Baker, Ryan S.J.d. (1989). *Data Mining for Education*. Carnegie Mellon University.

Baradwaj B. y Pal S. (2011). *Mining Educational Data to Analyze Student Performance.* International Journal of Advanced Computer Science and Aplication.

Bhattacharyya Saptashwa (2019). *Understanding Decision Tree Classification with Scikit-Learn*. Towards Data Science. Link: https://towardsdatascience.com/understanding-decision-tree-classification-with-scikit-learn-2ddf272731bd

Breiman Leo (2001), *Random Forest*. Statistics Deparment University of California.

Broome Kate (2018), *Who invented school?. Science Trends*. Link: https://sciencetrends.com/invented-school-created-standardized-education/

Brunsdon C., Charlton M. y Fortheringham A. (2002), *Geographically Weighted Regression: The analysis of spatially varying relationships*. University of Newcastle.

Centro de Implementación de Políticas Públicas para la Equidad y el Crecimiento (2017), *Aprender 2017, qué nos dicen los resultados*. Centro de Implementación de Políticas Públicas para la Equidad y el Crecimiento (Cippec). Link: https://www.cippec.org/textual/aprender-2017-que-nos-dicen-los-resultados

Chen T. y Gestri C. (2016), *XGBoost a Scalable Tree Boosting System*. University of Washington.

Clemens, Malbernat, Urrizaga y Varela (2015), *Aplicación de técnicas de Data Mining en Gestión de Docentes de Educación Superior. Impacto en el Desarrollo de la Profesión Académica*. Universidad CAECE.

Datos Argentina (2020), *Repositorio de Datos Públicos*. Datos Argentina. Link: https://datos.gob.ar/dataset/ign-unidades-territoriales/archivo/ign_01.04.01

Ente Nacional de Comunicaciones (2020), *Penetración por hogares nacional de Internet Fijo*. Datos Abiertos Ente Nacional de Comunicaciones. Link: https://datosabiertos.enacom.gob.ar/visualizations/29883/penetracion-por-hogares-nacional-de-internet-fijo/

Fuego Simondet, Javiar (2019), *La escuela, ante el desafío de enseñar para un mundo nuevo*. Diario La Nación. Link: https://www.lanacion.com.ar/opinion/educacion-la-escuela-ante-el-desafio-de-ensenar-para-un-mundo-nuevonota-de-tapa-nid2228662

Fundación Universidad Católica Argentina (2020). *La Educación de los Argentinos en Clave de Recupero y en Búsqueda de Oportunidades*. Universidad Católica Argentina.

Garabito y Vezub (2017), *Los profesores frente a la nueva/vieja escuela secundaria argentina*. Scielo. Link: http://www.scielo.org.mx/scielo.php?pid=S1607-40412017000100123&script=sci_arttext

Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy y Usama Fayyad (1996), *Advances in Knowledge Discovery and Data Mining*. IEEE Expert.

Gobierno de la Ciudad de Buenos Aires (2020), *¿Qué es AMBA?*. Buenos Aires Ciudad. Link: https://www.buenosaires.gob.ar/gobierno/unidades%20de%20proyectos%20especiales%20y%20puerto/que-es-amba

Llach J. y Cornejo M. (2016), *Factores condicionantes al aprendizaje en primaria y secundaria*. Secretaría de Evaluación Educativa del Ministro de Educación, Cultura, Ciencia y Tecnología de la Nación.

Molnar, Christoph (2020), *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Github. Link: https://christophm.github.io/interpretable-ml-book/

Naciones Unidas (2020), *Derechos Humanos*. Naciones Unidas Link: https://www.un.org/es/sections/issues-depth/human-rights/index.html

New Zealand Government (2015), *New Zealand education in 2025. Lifelong learners in a connected world*. New Zealand Ministry of Education. Link: https://www.education.govt.nz/assets/Documents/Ministry/Initiatives/Lifelonglearners.pdf

Peña Daniel (2002), *Análisis de Datos Multivariantes*. McGraw-Hill.

Prabha Lakshmi, S. (2014), *Educational Data Mining Applications*. Department of Computer Science, Seethalakshmi Ramaswami Collage.

Rivas, Axel (2010), *¿Quién controla el futuro de la educación?* Siglo XXI.

Rivas, Axel (2015), *América Latina después de PISA : lecciones aprendidas de la educación en siete países*. Fundación Centro de Implementación de Políticas Públicas para la Equidad y el Crecimiento (CIPPEC).

Romero C. y Ventura S. (2010), *Educational Data Mining: A Review of the State of the Art*. Universidad de Córdoba.

Senado y Cámara de Diputados Nacional (2014), *Ley 27.045*. Argentina Nación. Link: https://www.argentina.gob.ar/normativa/nacional/ley-27045-2014-240450

Templado, Ivana (2018), *Pruebas Aprender: o acerca de la cuantificación de oportunidades*. Fundación de Investigaciones Económicas Latinoamericanas.