EL DERECHO A LA EDUCACIÓN EN LA ARGENTINA Y LOS MODELOS DE PREDICCIÓN DE CIENCIA DE DATOS.

TRABAJO FINAL INTEGRADOR

ESPECIALIZACIÓN EN CIENCIA DE DATOS

INSTITUTO TECNOLÓGICO DE BUENOS AIRES
Tutora: Juliana Gambini

MARIA DANIELA RAFFO MASTRICOLA
Lic. en Administración de Empresas

Índice

1. INTRODUCCIÓN.

La educación es uno de los derechos humanos fundamentales. En la Argentina, si bien este derecho está legislado e institucionalizado, la desigualdad con la que se distribuye este servicio público es significativa. No hay lugar a la duda en que la administración del mismo ha sufrido numerosos cambios en las últimas décadas pero es transversal a cualquier época del país la desigualdad presente en la realidad de este derecho. Numerosas investigaciones demuestran que la calidad de la educación que reciben los diferentes alumnos del país difiere según la región en la que viven y el nivel socioeconómico que poseen.

El presente trabajo tiene como fin la presentación de diferentes técnicas de análisis de datos y modelos de predicción dentro de la familia de los árboles de decisión que indicarán de manera objetiva patrones regionales, socioeconómicos y culturales que afectan el nivel de educación de los jóvenes, como así también los factores más específicos que producen estos resultados. Se busca conocer aquellas variables particulares que puedan afectar el rendimiento del alumnado más allá de la provincia en la que viven o si asisten a una escuela pública o privada.

De la misma manera, este documento busca seguir alentando la utilización de modelos de ciencia de datos y técnicas de programación en las investigaciones sociales.

2. ESTADO DE LA CUESTIÓN.

En esta sección se abordan los conceptos principales que integran esta investigación, así también un breve repaso teórico de la evolución del derecho a la educación en Argentina en los últimos 30 años, mencionando lo acontecido en el año 2020 debido a la pandemia por Coronavirus. También se desarrolla el estado del arte de la utilización de la ciencia de datos en estudios de este ámbito.

a. Derecho a la educación en Argentina.

Para hablar de educación de calidad, primero se debe definir lo que es la calidad y lo que es el derecho a la educación. La calidad como concepto puede tener diversas acepciones como: que satisface un estándar de excelencia, que refiere a un propio valor, que cumple con un conjunto de atributos, entre otras. Por su parte, la educación en Argentina está legislada por la ley nacional 26026/2006, se la considera un bien público y este derecho debe ser garantizado por el Estado (Giuffré & Ratto, 2013).

El artículo 4 de esta ley introduce el concepto de educación de calidad cómo aquella educación integral, constante, que se debe impartir de forma equitativa y justa a todos los habitantes de la nación en conjunto con las organizaciones sociales y las familias (Honorable Congreso de la Nación Argentina, 2006).

b. Evolución de la educación en Argentina desde el regreso de la democracia en 1983.

Es importante que se considere la evolución que ha tenido este derecho en el país en los últimos 30 años. En la década de los 80 debido a las transformaciones globales que acontecieron, los países de América Latina sufrieron grandes crisis financieras. Argentina específicamente se encontraba en un proceso de democratización y transformación económica. Durante esta década lo que caracterizaba a la educación eran, por un lado, los presupuestos insuficientes, pero por otro el aumento de demanda de este servicio. También la pérdida del monopolio estatal de las instituciones escolares, y a la vez, un excesivo centralismo de los sistemas regulatorios.

Llegada la década de los 90 el diagnóstico que impulsaba una transformación educativa era la falta de equidad en la distribución de los servicios. Las reformas alentadas incluían: priorizar el derecho a la educación como medida política, desarrollar propuestas que favorezcan a los sectores con menos recursos, incrementar el presupuesto educativo, mejorar el

rendimiento de los alumnos, fortalecer la dirección de las instituciones educativas con el objetivo de mejorar su capacidad de gestión y autonomía, y así también fortalecer la relación de estas con sus entornos. (Centeno & Leal, 2011).

c. Explicación del proceso de privatización de la educación en Argentina.

En cuánto a los motivos del proceso de privatización de la educación argentina se pueden mencionar varias razones. En primer lugar, hay que destacar la relación entre esta elección y la calidad de las escuelas. Buena parte de los estudios señalan como justificación de esta decisión la búsqueda de mejores resultados académicos de los niños y adolescentes por parte de los padres. Otro motivo, son los maestros ausentes o escuelas cerradas que son mucho más comunes en el ámbito público. Para muchos padres era crucial el cumplimiento del calendario escolar no solo por la educación de sus hijos, sino también por la previsibilidad y la posibilidad de estos de concurrir a sus trabajos sin alteraciones. Relacionado con esto cabe mencionar la idea circular de un sistema educativo fuertemente segregado donde este mismo es solo un reflejo de la segregación social de la sociedad. Décadas atrás una 'élite' social intentaba monopolizar ciertas escuelas públicas impidiendo el ingreso de alumnos de otro nivel socioeconómico. Hoy en día, esa diferenciación se ve más entre las escuelas privadas, reservadas para familias con determinado ingreso. Otro de los motivos, siendo este punto de contradicción entre los autores, es el fenómeno de neoliberalismo y posneoliberalismo. Algunos autores relacionan estos procesos con el de privatización de la educación. La acción neoliberal, en general, ha impactado en el ámbito educativo con la privatización de instituciones, y fomentando la conexión entre el financiamiento y la gestión de estas de manera auto-suficiente. Para el caso argentino, se registraron más escuelas privadas en el periodo post-neoliberal (2002-2014).

Es importante señalar que también se dio un proceso de estatización de la educación privada en Argentina. El estado ha reconocido estas instituciones

diferentes acciones tales como la sanción de un régimen de subvenciones, la sanción de la Ley Federal de Educación de 1993 y su convalidación a través de la Ley Nacional de Educación de 2006. Hubo un reconocimiento del subsistema de la educación privada como integrante del universo educacional en la Argentina. Pero sí es crucial marcar la diferencia en cuanto a la mayor autonomía que presentan las escuelas privadas frente a las públicas.

Por último, hay autores que consideran que este proceso, en parte, estuvo incentivado por el propio Estado. Cómo se mencionó anteriormente, se denota debido al establecimiento de un sistema de subsidios a las escuelas privadas y la sanción de un marco legislativo regulatorio para estas instituciones. Para el estado, el crecimiento de la oferta educacional también por agentes privados, configuran una dinámica articulada que permite maximizar el financiamiento público por estudiante. (Narodowski et al., 2017).

d. El sistema educativo argentino en la actualidad.

Según datos del año 2020, el porcentaje de promoción efectiva, que refiere a la cantidad de alumnos que se matriculan en el siguiente año escolar, para la primaria era del 98,99%. Mientras que la tasa promedio del país para el secundario era del 93,24%. La provincia con peor índice de promoción efectiva para el primario era Corrientes, mientras que para el secundario era Entre Ríos. En ese mismo año, 0,68% de los alumnos primarios se matricularon como alumnos repitientes. Mientras que para el secundario ese porcentaje aumenta al 1,79%. Siendo nuevamente las provincias con mayor tasa de alumnos que repiten, Corrientes para el nivel primario y Entre Ríos para el nivel secundario. En cuanto al abandono escolar el promedio en la primaria es de 0,34% mientras que ese promedio en el país asciende al 4,97% para el secundario. El último año de educación obligatoria en el país tiene un índice de abandono interanual del 15,94%. (Dirección de Información Educativa, 2021).

En cuanto a los resultados de las últimas pruebas APRENDER llevadas a cabo en el país, en el año 2021 a alumnos de sexto de primaria, 45,2% de los

estudiantes evaluados obtuvieron un porcentaje acorde a una comprensión básica o por debajo del básico en matemáticas. Esta cifra es acorde a la obtenida en años anteriores. Si consideramos estudiantes solo de escuelas públicas el porcentaje de alumnos con desempeño deficiente en matemáticas es del 51,2% mientras que el porcentaje para alumnos de escuelas privadas es del 28,3%. En cuanto a la comprensión de conceptos de Lengua, 45% de los estudiantes evaluados presentan un rendimiento por debajo del satisfactorio. Siendo 51,4% y 23,2% el porcentaje de alumnos en escuelas públicas y privadas respectivamente con un resultado debajo de las expectativas. (Ministerio de Educación de la Nación, 2022)

Si hay una palabra que puede definir al sistema educativo en América Latina y específicamente en Argentina, esa palabra es: desigualdad. Tanto en los primeros años de escuela primaria como en los años de escuela secundaria se ve una gran diferencia en la calidad educativa según el nivel económico de los niños. Esta diferencia en la calidad y en los resultados académicos suele coincidir con la separación entre las escuelas públicas y privadas.

Estas diferencias se basan en si asisten a la escuela con doble escolaridad, si hablan lenguas extranjeras, su nivel de comprensión de los contenidos y sus habilidades con herramientas informáticas. A pesar de la heterogeneidad de cada sistema, para alumnos del primario que asisten a escuelas públicas su nivel de conocimiento en alguna lengua extranjera es del 20.1%, mientras que para aquellos que asistían a instituciones privadas este porcentaje aumenta al 45%. Como se mencionó anteriormente, también se puede observar una diferencia significativa en la compresión de contenidos de Lenguas, Matemáticas, Ciencias Sociales y Naturales. Para las escuelas privadas, el porcentaje de niños con compresión menor a básica o básica es reducido. Mientras que para aquellos alumnos que asisten a escuelas públicas ese porcentaje es el doble que en escuelas de gestión privada (Piovani, 2022).

e. La Educación y la pandemia por COVID-19.

La pandemia por Coronavirus provocó una crisis que ha atravesado todos los ámbitos humanos, y la educación no fue la excepción. La emergencia sanitaria ha dado lugar a la suspensión de las actividades educativas presenciales en más de 190 países. Según informó la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) para mayo del año 2020 alrededor de 1200 millones de alumnos en todo el mundo habían dejado de tener clases presenciales. Este cese de la actividad en vivo dio lugar al lanzamiento de modalidades de aprendizaje de forma virtual, utilizando diversas plataformas y herramientas (Comisión Económica para América Latina y el Caribe, 2020).

En una encuesta rápida realizada por UNICEF en octubre-noviembre del 2020, se obtuvo el dato que 93% de los hogares donde había niños y adolescentes se pudo mantener alguna actividad escolar. Sin embargo, las experiencias fueron muy diferentes y estuvieron condicionadas por los recursos del hogar y de las instituciones educativas. (Naciones Unidas En Argentina, 2021).

Asimismo, otro dato notorio, para Julio del 2020, solo 39% de los directores de escuelas secundarias afirmaron que podían seguir el plan educativo, siendo un 58% el porcentaje para los directores de educación primaria. En el caso de los adolescentes entre 13 y 17 años, cabe mencionar que un 18% de ellos no tenía acceso a internet en sus hogares iniciada la pandemia. Así también, el 37% de ellos manifestó no tener un dispositivo electrónico propio para su aprendizaje. Para octubre-noviembre del 2020, 12% de los adolescentes se sentían deprimidos y el 24% angustiados. El 65% afirmó que la pandemia afectó sus hábitos diarios y era difícil no hacer las actividades que solían hacer antes de las restricciones sanitarias (Naciones Unidas En Argentina, 2021).

f. Investigaciones de Ciencias Sociales y el uso de la Ciencia de datos.

El término Ciencias Sociales Computacionales es un campo interdisciplinario que se encuentra en la intersección de la Ciencia de datos y las Ciencias Sociales, que su principal foco son las inferencias causales y predictivas.

A medida que cada vez más las herramientas y medios digitales invaden las esferas humanas, hay cada vez más datos sobre los comportamientos humanos, y estos son de menester interés de análisis para las ciencias sociales. Todos éstos datos son digitales y es necesario contar con las herramientas computacionales adecuadas para captarlos, administrarlos, procesarlos y analizarlos.

Engel et al subrayan que existe un espacio en común entre los métodos estadísticos que se utilizan en el aprendizaje automático y los análisis causales y descriptivos de las investigaciones tradicionales de las ciencias sociales. Tales como, el análisis de los componentes principales (PCA), métodos de clasificación y de armado de grupos, también la regresión lineal y logística. Ambos campos de estudio buscan determinar conclusiones objetivas al hacer inferencias de datos no experimentales.

Estos mismos autores continúan señalando que las fortalezas de la minería de datos es la flexibilidad de los modelos que se pueden preparar, se pueden manejar relaciones no-lineales y las técnicas de validación de los resultados. Por otro lado, las fortalezas de las técnicas tradicionales de investigaciones sociales son el esfuerzo aplicado en el modelado de las variables latentes, también las técnicas estadísticas desarrolladas para manejar potenciales errores, cómo el factor de confusión. Son dos campos de Ciencias que pueden complementarse y fortalecerse mutuamente. (Engel et al., 2022).

g. La ciencia de datos y las investigaciones sobre la educación.

Continuando con lo anterior, si bien la estadística y las herramientas de ciencia de datos son disciplinas bien utilizadas hace años para clasificar, inferir y cuantificar la probabilidad de un fenómeno y ayudar en la toma de decisiones del mundo científico, su aplicación en áreas sociales y específicamente en el ámbito de la educación es más reciente.

Cómo señala Chan et al en el artículo "Aplicación de técnicas estadísticas multivariadas con el lenguaje de programación R en investigaciones educativas del nivel superior", el avance de estas tecnologías ha atravesado también al ámbito educativo y la incorporación de estos procedimientos han delineado un cambio en las investigaciones referentes a este tema. Cabe mencionar que más allá de que el uso de software ahorra tiempo y permite un trabajo mucho más flexible y dinámico, estas herramientas no son suficientes para una comprensión integral de los fenómenos. En este contexto, ocupa un rol protagónico el dominio de la lógica estadística aplicada y la correcta interpretación de los resultados para poder entender como válidas las conclusiones. (Chan et al., 2020).

En la actualidad, se cuenta con mucha potencia para la captura de datos digitales, esto provoca que las investigaciones sean de mayor confianza dado la representatividad de las muestras que se pueden obtener. En este contexto, Zengin et al, autores del artículo " Un estudio de muestra sobre la aplicación de técnicas de investigación de minería de datos en la ciencia de la educación: el desarrollo de un mayor significado de los datos", ejemplifican cómo la ciencia de datos ha ganado popularidad entre los investigadores. Estás técnicas no son una solución en sí mismas si no que son herramientas que apoyan el proceso de toma de decisión, ayudan a encontrar patrones y relaciones entre los datos. Nuevamente, se vuelve a afirmar que el uso de la ciencia de datos en las ciencias sociales, y más que nada, en la educación, es muy reciente. La mayoría de estos estudios son llevados a cabo por profesionales de las ciencias computacionales o gerentes pero no son preparados por profesionales del ámbito de la educación. (Zengin et at., 2011)

En cuanto a los trabajos sobre educación, los softwares más utilizados son: SPSS (Statistical Package for the Social Sciences), STATA (Statistical Software for Data Science), SAS (Statistical Analysis Software), entre otros. Sin embargo, muchos de estos softwares se trabajan bajo licencias pagas, no resultan tan flexibles ni la interfaz es tan amigable para el usuario. Asimismo, también existen software de licencias libres que se emplean para este tipo de investigaciones, el más utilizado en los últimos años, y cada vez con cada vez mayor preponderancia es R. Para mayor información sobre los programas ver SPSS IBM Corp., STATA STATA Corp., SAS SAS institute inc. y R R Core Team, 2023.

Cómo mencionan los autores Chan et al, e l lenguaje de programación de alto nivel R ha incursionado profundamente en los estudios del ámbito educativo, se ha evidenciado un crecimiento sostenido de su uso, con un pico en las publicaciones en el año 2018. Después de este año se nota un leve descenso en el uso de este software debido a que la comunidad científica ha optado por otras herramientas cómo es Python. No obstante, en el área de Ciencias Sociales específicamente se muestra una leve tendencia a la preferencia de R sobre Python. (Chan et al., 2020).

También se puede mencionar el uso de SQL server en estudios de ciencia de datos sobre educación. Se pueden analizar los datos usando árboles de decisión, redes de dependencia y técnicas de clustering sobre SQL server. (Zengin et at., 2011).

Este contexto, plantea nuevos desafíos para los profesionales de las Ciencias Sociales, es menester el desarrollo de habilidades de programación para poder acceder a todos los beneficios del uso de estas técnicas y herramientas (Chan et al., 2020).

3. DEFINICIÓN DEL PROBLEMA.

La tasa de comprensión de lengua y matemática en la Argentina varía significativamente entre regiones, así como también por el nivel socioeconómico y el nivel cultural del hogar de los alumnos. Sin embargo, no se conocen cuáles son los factores específicos que generan tales diferencias.

4. JUSTIFICACIÓN DEL ESTUDIO.

Si bien ya se han realizado estudios que denotan y señalan las diferencias en la distribución del servicio de la educación en el país, hay una carencia en la identificación de que factores más concretos llevan a estos resultados. Se conoce que existe una inequidad según la región en la que vive el alumno y su nivel socioeconómico, pero se desconoce si hay variables particulares tales como poseer o no una computadora propia, distancia al establecimiento educativo, relación con los compañeros en la institución, entre otros, que puedan ser los que finalmente tengan mayor preponderancia en los resultados académicos de los alumnos. Esta investigación tiene por objetivo encontrar patrones para salvaguardar ese vacío de conocimiento.

Asimismo, con el mismo, se busca seguir incursionando en la utilización de herramientas de software en el ámbito social, y específicamente en investigaciones sobre educación.

5. ALCANCE DEL TRABAJO Y LIMITACIONES.

Este trabajo tiene como interesados a aquellas partes involucradas en el área de la educación cómo docentes, funcionarios del gobierno que trabajan en educación y a cualquier otro interesado en la materia. También podría ser de interés para aquellos profesionales de las ciencias sociales que desean incursionar en el uso de las técnicas de Ciencia de datos para optimizar sus trabajos de investigación.

El presente documento utiliza cómo fuente de datos las pruebas Aprender que lleva a cabo el ministerio de Educación. El sistema de evaluación alcanza a estudiantes de todo el país, tanto del nivel primario cómo del secundario. El mismo

está enfocado en la comprensión de conceptos de Lengua y Matemáticas, dejando afuera la comprensión de contenidos de otras materias. Junto con las evaluaciones se realiza una encuesta de índole sociodemográfica y de las condiciones con las que cuenta el alumno para su desarrollo intelectual. Cabe mencionar que hay ciclos en los que se evalúa a las escuelas primarias y secundarias, hay otros ciclos en los que solo se evalúa a alumnos del nivel primario y otros en los que participan solo estudiantes del secundario. Para el presente trabajo se utilizan las bases de las pruebas Aprender del año 2021, las mismas fueron realizadas a alumnos de sexto de primaria.

También se señala que en el año 2020 debido a la pandemia por Covid-19 no se realizaron las pruebas APRENDER en el país, retomando las mismas en el año 2021. Los resultados de las pruebas llevadas a cabo en el año 2022 no se encontraban disponibles al momento de comenzado el presente trabajo.

6. HIPÓTESIS.

Con los datos de las pruebas APRENDER sobre la edad, sexo, conformación del hogar, servicios con los que cuenta el alumno en su domicilio, actividades del alumno fuera de la escuela, calidad del servicio que se brinda en el establecimiento educativo, entre otros factores, se pueden identificar patrones sociales, económicos, culturales que afectan el índice de comprensión de Lengua y Matemática. Además, se puede conocer que factores específicos de los que contiene la base de datos tienen una mayor preponderancia en el rendimiento del alumnado.

a. Variables Contextuales.

- Provincia en la que vive el alumno.

b. Variables Independientes.

- Edad.
- Sexo.
- Personas con las que vive el alumno.
- Nivel educativo de las personas con las que vive el alumno.
- Servicios con los que cuenta el alumno en su hogar.

- El alumno realiza alguna clase de trabajo además de estudiar.
- Nivel de atención de la maestra al alumno.
- Si en la escuela sufre alguna clase de discriminación.

c. Variables Dependientes.

- Desempeño en lengua.
- Desempeño en matemática.

7. OBJETIVOS.

a. Objetivo general:

Identificar patrones regionales, socio-económicos y culturales que puedan explicar el desempeño de los alumnos en cuanto a la comprensión de lengua y matemática. Pero además, identificar qué factores específicos tienen un mayor peso en la obtención de estos resultados.

b. Objetivos específicos:

- Construir la base de datos con los datos de las pruebas APRENDER del año 2021. Limpiar y consolidar estos datos con el fin de no tener datos faltantes, ni inconsistencias en cuanto a formato.

- Analizar estos datos con gráficos e indicadores descriptivos para tener una primera identificación de los patrones más básicos cómo aquellos referidos a la provincia, sector o ámbito de los alumnos.

- Construir al menos dos modelos de predicción que puedan clasificar a los estudiantes según el rendimiento en las pruebas, así también que estos modelos puedan identificar los factores que tienen mayor peso en la obtención de esos resultados.

- Comparar ambos modelos en cuanto a su rendimiento y resultados.

8. METODOLOGÍA.

La metodología utilizada para llevar adelante el siguiente trabajo es una metodología ágil específica del campo de la ciencia de datos, KDD (Knowledge Discovery from Data). La misma consiste en cuatro etapas principales: integrar y recopilar los datos, aplicar las técnicas de limpieza y transformación de datos necesarias, diseñar los modelos de predicción, y la cuarta y última etapa, que es la de análisis e interpretación de los resultados.

a. Técnicas.

En primer lugar, se usan técnicas de consolidación de base de datos para lo que es la integración de las pruebas. También se utilizan técnicas de limpieza y transformación de los datos para que presenten el mismo formato y no haya inconsistencias. En segundo lugar, a partir de fórmulas y gráficos de estadística descriptiva se realiza un primer análisis de la base de datos.

Posteriormente, es menester la separación de los datos de forma aleatoria entre un conjunto de datos para entrenar y otro para testear. Por último, se diseñan modelos de predicción, se utilizan modelos de árbol de decisión, modelos de Random Forest y modelos de Extreme Gradient Boosting. Es importante la realización de al menos dos modelos diferentes para poder comparar los resultados entre estos y que las conclusiones sean más integrales.

b. Herramientas.

Las herramientas para llevar adelante esta investigación son, en primer lugar, SQL para el armado de la base de datos de forma integral. Para el análisis de los datos, su limpieza y transformación se utiliza el lenguaje de programación de alto nivel R. Asimismo, los modelos de predicción se realizan y comparan en R.

9. ANÁLISIS EXPLORATORIO DE LOS DATOS:

a. Base de datos.

Cómo se mencionó previamente, para este trabajo se utiliza la base de datos de alumnos de las pruebas Aprender del año 2021. Se pueden encontrar las mismas en la página del Ministerio de Educación, en el sector de Evaluación e información educativa, seleccionado el año 2021. Se puede acceder a las bases por el siguiente link: https://www.argentina.gob.ar/educacion/evaluacion-informacion-educativa/aprender/aprender-2021.

Para poder realizar la carga y el armado de la base de datos en SQL, se reemplaza los valores en blanco por un valor nulo para hacer referencia que en esos campos no había un dato real. La base de datos pública cuenta con 40207 registros sin valores (con espacio en blancos), lo que representa un 6,32% de la base total.

Se crea la base de datos "Pruebas_Aprender" para realizar un primer análisis de la misma. Cuenta con la tabla Alumnos, que contiene 635515 registros y 147 variables. También se crean tablas anexas, Jurisdicción, Ámbito y Sector, para una mejor visualización de los resultados.

b. Análisis descriptivo de la base de datos.

i. Distribución por provincia de los alumnos.

Se puede observar en la Tabla 1 la distribución de los estudiantes encuestados. Se observa que las provincias que tienen mayor representación en la base de datos son Buenos Aires, Córdoba y Santa Fe, que coincide con las que poseen mayor cantidad de habitantes.

Tabla 1. Distribución de alumnos por provincias.

Provincia	Cantidad de alumnos
BS. AS.	239584
Córdoba	52366
Santa Fe	46957
C. A. Bs. As.	33460
Mendoza	28010
Tucumán	23329
Misiones	23063
Salta	22657
Entre Ríos	19898
Corrientes	19234
Chaco	18963
Santiago del Estero	17889
Jujuy	12411
San Juan	11855
Rio negro	10506
Formosa	10112
Chubut	8551
San Luis	7492
Catamarca	6466
La Rioja	5979
La Pampa	5178
Santa Cruz	4962
Neuquén	4024
Tierra del Fuego	2569

ii. Distribución por ámbito de los alumnos.

Las pruebas se realizan tanto en escuelas de zona urbana como de zona rural. Se observa en la Tabla 2 que hay más alumnos evaluados en zonas urbanas dado que actualmente las personas que viven en zonas rurales a nivel país son la minoría.

Tabla 2. Distribución de alumnos por ámbito.

Ámbito	Cantidad de estudiantes
Rural	61343
Urbano	574172

iii. Distribución por sector de los alumnos.

Continuando con lo anterior, se evalúa a alumnos tanto de escuelas públicas, como privadas. La distribución se puede ver en la Tabla 3, el volumen de alumnos evaluados es más significativo en las escuelas estatales.

Tabla 3. Distribución de alumnos por sector.

Sector	Cantidad de estudiantes
Estatal	466084
Privado	169431

iv. Nivel de desempeño en Lengua y en Matemática por provincia.

De la misma manera, se realiza un primer análisis de correlatividad entre algunas variables seleccionadas con el fin de obtener un primer acercamiento a patrones. Teniendo en cuenta aquellos alumnos que no tienen valores nulos para la puntuación en las evaluaciones de Matemáticas y Lengua, se calcularon promedios por provincia.

Se puede observar en la Tabla 4 que la provincia que presenta un peor desempeño en ambas esferas es Chaco. A su vez, el sector que presenta mejor desempeño en el año 2021 es la Ciudad Autónoma de Bs. As.

Podemos observar que el máximo en promedio que presenta una provincia es 3 que equivale a un desempeño satisfactorio. El resto de las provincias están en un número más cercano a 2 que significa un nivel básico. Asimismo se observa, que en promedio, los alumnos obtienen mejores resultados en Lengua que en Matemática.

Tabla 4. Desempeño en Lengua y en Matemática por provincia.

Provincia	Desempeño en Lengua	Provincia	Desempeño en Matemática
Chaco	2.21	Chaco	2.16
Catamarca	2.30	Catamarca	2.16
Santiago del Estero	2.32	La Rioja	2.20
San Juan	2.36	Misiones	2.29
Corrientes	2.37	Corrientes	2.29
Misiones	2.38	Entre Ríos	2.31
Entre Ríos	2.42	San Juan	2.33
La Rioja	2.44	Santa Cruz	2.33
Salta	2.48	San Luis	2.37
Tucumán	2.48	Santiago del Estero	2.39
San Luis	2.48	Neuquén	2.41
Jujuy	2.51	Chubut	2.44
Formosa	2.54	Jujuy	2.46
Neuquén	2.55	Salta	2.46
Santa Fe	2.57	Tierra del Fuego	2.46
Santa Cruz	2.57	Tucumán	2.46
Mendoza	2.60	BS. AS.	2.47
BS. AS.	2.60	Santa Fe	2.49
Rio negro	2.61	Formosa	2.50
Chubut	2.63	Rio negro	2.53
La Pampa	2.69	Mendoza	2.59
Tierra del Fuego	2.75	La Pampa	2.61
Córdoba	2.82	Córdoba	2.80
C. A. Bs. As.	3.03	C. A. Bs. As.	2.90

v. Nivel de desempeño en Lengua y en Matemática por ámbito.

Se observa en la Tabla 5 el desempeño promedio por ámbito, sin tener en cuenta los valores nulos. Se destaca que el desempeño en lengua es más bajo en el ámbito rural pero es más alto en matemáticas respecto a las escuelas urbanas.

Tabla 5. Desempeño en Lengua y en Matemática por ámbito.

Ámbito	Desempeño en Lengua	Desempeño en Matemática
Rural	2.43	2.54
Urbano	2.61	2.49

vi. Nivel de desempeño en Lengua y en Matemática por sector.

Así también, si se analiza el desempeño promedio por sector, se observa en la Tabla 6, que el desempeño es más bajo en matemáticas que en lenguas en general. En este caso se señala que el desempeño en los colegios públicos es notablemente más bajo en promedio que en los colegios privados para el año 2021.

Tabla 6. Desempeño en Lengua y en Matemática por sector.

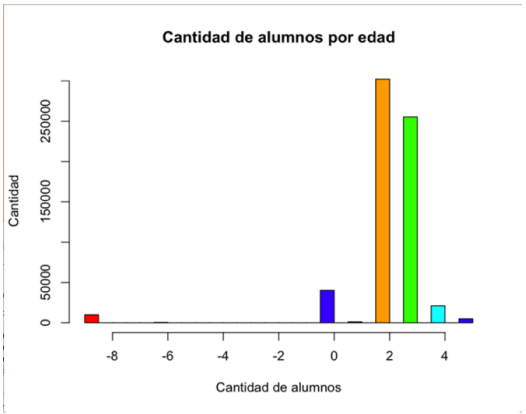
Sector	Desempeño en Lengua	Desempeño en Matemática
Estatal	2.40	2.35
Privado	3.08	2.86

vii. Rango etario.

En la Figura 7 se encuentra la distribución absoluta de los estudiantes encuestados por edad. En la edición del año 2021 de las "Pruebas Aprender" se evaluó a estudiantes de sexto grado de primaria.

En la Figura 7, los valores negativos refieren a estudiantes que completaron con varios valores a la vez. Los valores 0 a aquellos que no completaron su edad. El 1 refiere a estudiantes avanzados de grado, de 10 años o menos. El 2 se refiere a estudiantes de 11 años, el 3 a aquellos de 12 años. La edad acorde para estar en sexto grado. El 4 y 5 refieren a estudiantes que por algún motivo se atrasaron en sus estudios primarios y tienen 13, 14 o más años.

Figura 7. Rango Etario.



10. MODELOS PREDICTIVOS:

a. Armado de modelos, conjunto de entrenamiento y testeo.

Los modelos predictivos utilizando técnicas de aprendizaje automático deben presentar un esquema de trabajo que cuente con las siguientes etapas:

I. Conocer a profundidad la problemática y sus diferentes aristas.
II. Disponer de los datos referentes a ese problema. Y realizar un análisis de los mismos.
III. Seleccionar los modelos predictivos propicios para esos datos.
IV. Modelar el caso.
V. Testear los modelos.
VI. Optimizar los parámetros de cada modelo para obtener el rendimiento óptimo posible.
VII. Comparar los diferentes modelos y seleccionar el óptimo.

Estos mismos pasos fueron respetados para el presente trabajo.

Así mismo se debe considerar que existen diferentes tipos de casos para modelar. Si las variables del caso son cuantitativas estamos frente a un modelo de regresión. Si las variables con las que se cuenta son cualitativas es un caso de clasificación.

En la base de datos trabajada las variables dependientes o a predecir son de tipo categóricas, ergo, estamos ante un caso de clasificación. La base de datos cuenta con cuatro categorías con las que se evalúa el desempeño de los alumnos, nivel por debajo del básico, nivel básico, nivel satisfactorio y nivel avanzado. Para trabajar los modelos de aprendizaje automático se agrupan estas cuatro categorías en dos: alumnos con un nivel no satisfactorio (aquellos con desempeño por debajo del básico y básico) y alumnos con nivel satisfactorio (aquellos con desempeño satisfactorio o avanzado). La razón de ésto radica en que con el presente trabajo nos interesa diferenciar aquellos factores específicos que no ayudan a que los alumnos obtengan resultados satisfactorios en sus evaluaciones de Lengua y Matemática. Asimismo,

dado que la base de datos presenta dos variables a predecir, el desempeño en Lengua, y el desempeño en Matemáticas, se prepara y se trabaja con modelos por separado para cada variable objetivo. De esta manera se pueden focalizar mejor los resultados para indicar los factores específicos que afectan el rendimiento de los estudiantes en cada una de estas áreas.

Es menester señalar que al trabajar con modelos de predicción es necesario separar a la base de datos en dos conjuntos, un conjunto de entrenamiento, y otro de testeo. Este punto es de crucial importancia, ya que de forma contraria, no se cuenta con datos para realizar la validación del funcionamiento del modelo ya que él mismo trabajó y aprendió con todos los datos de la base.

Cómo se mencionó anteriormente, para realizar los modelos se prescinde de 40207 registros ya que cuentan con algún dato faltante, para que estos registros no afecten la performance de los modelos. Eliminar los mismos no perjudica la integridad del análisis dado que estos representan solamente el 6,32% del total de la base de datos. Asimismo se deja de lado ciertas columnas, tales cómo "Id de los alumnos" y de columnas de ponderación para no cargar al modelo con columnas sin valor agregado. Posteriormente, se divide la base, un 75% de los datos quedan en el grupo de entrenamiento y un 25% de los mismos en el conjunto de testeo.

b. Explicación teórica de los modelos seleccionados.

Para cumplir con los objetivos planteados en el presente trabajo, se utilizan tres algoritmos diferentes dentro de la familia de los árboles de decisión. Se inicia con modelos de árboles simples, se continúa con modelos de Random Forest, y por último se experimenta con modelos de Extreme Gradiente Boosting. Cada uno de estos son versiones mejoradas del anterior y esto se puede observar en los resultados de performance de los modelos.

I. Primer Modelo: Árbol de decisión.

El primer modelo que se utiliza es un árbol de decisión, un algoritmo dentro de la categoría de aprendizaje supervisado no paramétrico. Este modelo se puede utilizar tanto para casos de regresión cómo de clasificación.

Cómo se expresó de forma anterior, el objetivo principal de este estudio es el análisis e identificación de patrones en el rendimiento del alumnado. Un árbol de decisión comienza con un nodo raíz, el cual alimenta nodos internos, estos a su vez presentan nodos hojas o terminales. El algoritmo identifica los cortes óptimos para la toma de decisiones. De esta manera se puede identificar cuáles de las variables de nuestra base de datos son los mejores cortes para el árbol, es decir, qué variables tienen mayor peso en el desempeño de los alumnos tanto en Lengua, cómo en Matemática.

Existen diferentes tipos de árboles que se pueden construir. Para este caso se utilizan los árboles de decisión de la librería "CART" del lenguaje de programación R. Estos se basan en el algoritmo de Hunt, así también utilizan la medida de división del índice de Gini para seleccionar el atributo de división óptimo. (Priyam et. Al, 2013).

El algoritmo de Hunt consiste en buscar el mejor corte para cada nodo, cada nodo tiene asociado un subconjunto de los datos de entrenamiento que de acuerdo al mejor corte encontrado se subdividen en nodos hijos. Así también, se indica que un nodo es puro, si el modelo le asigna a todos los registros del mismo su correcta clasificación. Para realizar esta medición se utiliza el índice de Gini:

$$\text{Gini (t)} = 1 - \sum_{j} [p(j \mid t)]^2$$

Siendo $p(j \mid t)$ la probabilidad de que un caso esté en la clase j siendo que se asignó al nodo t. (Haro Rivera, 2020).

De igual manera, es menester señalar los principales parámetros de este modelo. Haciendo referencia a los más significativos, "Minsplit",

refiere al número mínimo de observaciones necesarias que debe haber en un nodo para que se efectúe un corte. En relación a este, El valor de "Minbucket" refiere a la cantidad de observaciones que deben restar en el nodo terminal de forma posterior a un corte. Otro parámetro crucial es el valor de "Cp" o parámetro de complejidad para la poda del árbol. Un valor igual a 1 de "Cp" expone a un árbol sin divisiones y un valor de 0 a un árbol de máxima profundidad. Además este valor sirve para indicarle al modelo la proporción de reducción de error mínima necesaria para realizar un corte. Valores de complejidad más cercanos a 1 simplifican el modelo y reducen el tiempo de cómputo. Asimismo, el valor de "Maxdepth" corresponde a la profundidad máxima del árbol.

II. Segundo modelo: Random Forest.

El segundo modelo que se ejecuta para este trabajo son modelos de Random Forest. Estos modelos son la combinación de un determinado número de árboles de predicción, cada uno de estos depende de un vector aleatorio independiente, pero para todos los árboles del bosque se utiliza la misma distribución. La generalización del error para los casos de clasificación depende de la fortaleza de cada árbol en particular y de la correlación entre los mismos. Este modelo también se podría utilizar para casos de regresión. (Breiman, 2001).

En un modelo de Random Forest, para cada árbol N del bosque, se genera un vector aleatorio Y, independiente de los vectores pasados pero con la misma distribución. La selección del vector consiste en un número aleatorio entre 1 y K. La naturaleza y dimensión del mismo depende de su uso en la construcción de los árboles. Se construye un árbol de decisión N usando el conjunto de entrenamiento y el vector Y, obteniendo cómo resultado un árbol de clasificación N(x, y), donde x es un vector de entrada. De este mismo modo, se construyen un

número determinado de árboles, donde cada uno castea un voto a la clase más popular de X. (Breiman, 2001).

Random Forest = {N (x, y), k = 1, . . .}

Los parámetros principales de este modelo son, el valor de "Ntree" el cuál indica el número de árboles que componen el bosque. Cabe señalar que la cantidad necesaria de árboles para una performance adecuada del modelo aumenta a medida que lo hacen las variables independientes. Sin embargo, hay que considerar que la cantidad de árboles del modelo es proporcional al tiempo de cómputo que requiere el mismo. Asimismo, el parámetro "Mtry" indica el número de variables independientes seleccionadas aleatoriamente para cada división de los nodos. De igual manera, "Nodesize" señala el tamaño mínimo de un nodo, si el mismo no cuenta con un número determinado de registros no se produce una nueva división. (Alaminos-Fernández, 2022)

Además, este mismo modelo produce adicionalmente dos elementos de información de interés: las variables de importancia y la medida de proximidad entre observaciones. Las variables de importancia, se estima la importancia de cada una de las variables al analizar cuánto aumenta el error de predicción si se permuta todos los valores de esa variables, cuando en paralelo se deja estables los valores del resto de las variables predictores. Se hace este cálculo para cada árbol a medida que se construye el modelo. La matriz de proximidad se refiere a la fracción de elementos i y j que finalizan en el mismo nodo. Es esperable que aquellos registros con valores similares queden agregados en el mismo nodo. Esto es de utilidad para identificar estructuras en la base de datos. (Liaw & Wiener, 2002).

III. Tercer modelo: Extreme Gradient Boosting o XGBoost.

El tercer modelo de aprendizaje automático que se desarrolla en este trabajo son modelos de "Extreme Gradient Boosting" o XGBoost. Este es un algoritmo de aprendizaje automático que combina la creación de modelos de predicción de manera secuencial y la técnica de Boosting (reducción del error en los modelos predictivos).

En primera instancia, es un modelo de predicción escalable construido de extremo a extremo, donde se desarrollan árboles que toman las técnicas de Boosting para reducir el error. Se crean árboles de forma secuencial y cada uno va aprendiendo de su antecesor, árboles más débiles crean otros posteriores más robustos.

Además, este algoritmo, incluye dos técnicas para evitar el Overfitting: la contracción o "Shrinkage" y el sub-muestreo de las columnas. El Overfitting es un concepto crucial en el campo de la Ciencia de datos. El mismo hace referencia a modelos que se sobre-ajustan a los datos de entrenamiento, y por ende posteriormente, no pueden realizar predicciones correctas sobre datos desconocidos. La primera técnica mencionada fue introducida inicialmente por Friedman. Las escalas de contracción añaden nuevos pesos por un factor N después de cada iteración. Esto reduce la influencia de cada árbol individual y deja espacio para futuros árboles para mejorar el modelo. La segunda técnica refiere a no usar todas las columnas de la base de datos en cada iteración. Esto no solo reduce el Overfitting si no que ayuda a mejorar la performance computacional del modelo.

Asimismo, este modelo está programado para trabajar con la dispersión que puede haber en los datos. Cuando falta un valor en la matriz, la instancia se clasifica en la dirección predeterminada, estas se aprenden de los mismos datos.

Siguiendo con el propósito de mejorar el tiempo computacional del modelo, este almacena los datos en unidades de memoria que se los conoce como “bloques”. Los datos de cada bloque se almacenan en el formato de columna comprimida (CSC), y cada columna es ordenada por el valor de característica correspondiente. Este diseño de datos de entrada solo necesita ser computado una vez antes del entrenamiento, y puede ser reutilizado en iteraciones posteriores.

Es debido a todo esto que hoy en día XGboost es uno de los algoritmos más utilizados en la Ciencia de datos. Se adapta a diversos problemas y diversas clases de bases de datos. El mismo se puede programar en diversos lenguajes de programación y es un paquete de código abierto. (Chen & Guestrin, 2016).

Los principales parámetros de este modelo son: la función objetivo donde se indica el tipo de modelo necesario para la variable a predecir (clasificación-regresión). Asimismo, se debe indicar la medida de evaluación del modelo, el parámetro de “Eval_metric”. Además, el parámetro “Nrounds”, él mismo señala la cantidad de iteraciones a realizar. Un número más elevado de iteraciones, por lo general, ayuda a mejorar la performance del modelo pero incrementa el tiempo de cómputo. El parámetro de “Maxdepth”, cómo se presentó para los modelos anteriores, indica la profundidad de cada árbol que se utiliza. El valor de “ETA” o “Learning Rate” indica la tasa de aprendizaje y contribución de cada árbol al modelo, un valor más pequeño conlleva a modelos más robustos pero se deben tomar medidas para evitar el Overfitting. Otro parámetro importante es el “Colsample_bytree”, el mismo le indica al modelo la cantidad de variables independientes a tomar para la construcción de los árboles.

c. Indicadores para medir la performance de los modelos.

Para medir la performance de los modelos, se utilizan diferentes indicadores. El primero de estos, es el índice de exactitud o "Accuracy", refiere a la proporción de predicciones correctas sobre el total de predicciones realizadas.

$$Accuracy = (VP + VN) / (VP+FP+VN+FN)$$

Siendo "VP", verdaderos positivos, "VN", verdaderos negativos, "FP", falsos positivos, "FN", falsos negativos.

También es importante analizar los indicadores de Recall y la tasa de precisión negativa o tasa de exclusión. El primer término refiere a los casos positivos correctamente detectados y el segundo a los casos negativos correctamente detectados. (Alaminos-Fernández, 2022)

$$Recall = VP / (VP+FN)$$

$$Precisión = VN / (VN + FP)$$

Para el caso de los modelos de Random Forest, otra medida que se puede utilizar es analizar la tasa de error "Out of Bag" o "OBB" del modelo. Cómo se mencionó anteriormente, un modelo de Random Forest, crea múltiples árboles de decisión utilizando subconjuntos de datos aleatorios. En este proceso, un porcentaje de los datos se excluyen y no se utilizan para ese árbol en particular. Una vez terminado el modelo, esos datos excluidos aleatoriamente, el modelo los utiliza para hacer predicciones y comparar con los valores reales. La tasa de error de "OBB" es la proporción de predicciones incorrectas sobre el total de las realizadas.

Por último, para comprobar que los modelos óptimos seleccionados no presentan sobreajuste o "Overfitting", se los utiliza para predecir el conjunto de entrenamiento (el mismo conjunto de datos del que los modelos aprendieron), y se corroboran los mismos

indicadores que para las predicciones sobre el conjunto de testeo. Si los indicadores de las predicciones sobre el conjunto de entrenamiento son significativamente más favorables que aquellas sobre el conjunto de testeo, nos indica que el modelo esta sobre-ajustado a los datos, ergo, no es un modelo de aprendizaje automático confiable.

d. Parámetros utilizados y resultados de los modelos.

i. Parámetros y resultados del primer modelo: Árbol de decisión.

Cómo se observa en las Tablas 9 y 10, se crearon diferentes versiones de modelos de árbol de decisión utilizando diferentes combinaciones de parámetros. Para encontrar el modelo óptimo se realiza una búsqueda en rejilla o "Grid Search". Para la variable objetivo de Lengua, para el parámetro de "MinSplit", se realiza una exploración entre el rango [50,400]. Mientras que para el parámetro de "MinBucket" se busca el óptimo entre el rango [10,250]. Para el caso de la variable objetivo de Matemática, también se realizan búsquedas en rejilla para estos dos parámetros. Siendo los rangos de [100,500] y de [50,300] para los parámetros de "MinSplit" y "MinBucket" respectivamente.

El modelo que presenta índices de rendimiento más favorables para el conjunto de Lengua es el número 18, para el cual se asignó un "Minsplit" de 150, un "Minbucket" de valor 100 y finalmente se designó un "CP" de 0.1. Con este modelo se obtiene un valor de exactitud de 0.7128, con un Recall de 0.5602 y una precisión de 0.8094. Si bien hay modelos que presentan un índice de exactitud con un pequeño incremento, estos mismos a la vez presentan peores índices de Recall, por eso el modelo elegido como el óptimo fue el número 18. Del mismo modo, se observa en la Tabla 10 el modelo con mejor performance para la variable de Matemáticas utiliza un "Minsplit" de 300 y un "MinBucket" de 220. Este modelo presenta un índice de exactitud levemente menor que el modelo de Lengua, el

mismo es de 0.6938. Además, valores de 0.5474 y 0.7948, de Recall y precisión respectivamente.

Tabla 9. Parámetros utilizados y resultados de los modelos la variable objetivo Lengua:

Nro. Modelo	CP	MIN_SPLIT	MIN_BUCKET	ACCURACY	Recall	Precisión	Variable Objetivo
1	0.1	50	10	0.6855	0.5449	0.7745	Lengua
2	0.1	50	20	0.6913	0.5483	0.7818	Lengua
3	0.1	50	30	0.6972	0.5489	0.7911	Lengua
4	0.1	50	40	0.7041	0.55	0.8015	Lengua
5	0.1	50	50	0.7078	0.5621	0.8	Lengua
6	0.1	50	60	0.7089	0.5597	0.8033	Lengua
7	0.1	50	100	0.7128	0.56	0.8094	Lengua
8	0.1	75	10	0.6944	0.5482	0.7870	Lengua
9	0.1	75	20	0.6981	0.5516	0.7907	Lengua
10	0.1	75	30	0.6996	0.5501	0.7942	Lengua
11	0.1	75	50	0.7078	0.5621	0.8	Lengua
12	0.1	75	100	0.7128	0.5602	0.8094	Lengua
13	0.1	75	150	0.7156	0.5579	0.8154	Lengua
17	0.1	150	50	0.7096	0.5608	0.8037	Lengua
18	0.1	150	100	0.7128	0.5602	0.8094	Lengua
19	0.1	150	200	0.7166	0.5473	0.8237	Lengua
20	0.1	200	250	0.7166	0.5481	0.8232	Lengua
21	0.1	400	250	0.7166	0.5481	0.8232	Lengua
22	0.1	150	120	0.7155	0.5585	0.8149	Lengua

Tabla 10. Parámetros utilizados y resultados de los modelos la variable objetivo Matemática:

Nro. Modelo	CP	MIN_SPLIT	MIN_BUCKET	ACCURACY	Recall	Precisión	Variable Objetivo
1	0.1	150	100	0.6885	0.5515	0.7831	Matemática
2	0.1	200	150	0.69	0.5558	0.7826	Matemática
3	0.1	300	200	0.6935	0.5567	0.7879	Matemática
4	0.1	500	300	0.6929	0.5477	0.7932	Matemática
5	0.1	300	150	0.69	0.558	0.7826	Matemática
6	0.1	100	50	0.6801	0.5535	0.7675	Matemática
7	0.1	400	250	0.6933	0.5472	0.7942	Matemática
8	0.1	400	300	0.6929	0.5477	0.7932	Matemática
9	0.1	300	220	0.6938	0.5474	0.7948	Matemática

Resultados de los modelos óptimos de árbol de decisión utilizando los conjuntos de entrenamiento:

Como se observa en la Tabla 11, los indicadores de performance de los modelos óptimos tanto para Lengua cómo para Matemática, utilizando los conjuntos de entrenamiento, presentan valores similares, levemente más favorables, que aquellos utilizando los conjuntos de testeo. Estos resultados confirman que los modelos no presentan sobreajuste a los datos con cuales aprendieron y se los puede considerar modelos de predicción correctos.

Tabla 11. Resultados de los modelos óptimos utilizando los conjuntos de entrenamiento:

					Conjunto de Testeo			Conjunto de Entrenamiento		
Variable Objetivo	Nro. Modelo	CP	MIN SPLIT	MIN BUCKET	Accuracy	Recall	Precisión	Accuracy	Recall	Precisión
Lengua	18	0.1	150	100	0.7128	0.5602	0.8094	0.7405	0.5944	0.8324
Mat.	9	0.1	300	220	0.6938	0.5474	0.7948	0.7074	0.5655	0.8053

ii. Parámetros y resultados del segundo modelo: Random Forest.

Cómo se observa en las Tablas 12 y 13, si bien los resultados de los modelos de Random Forest fueron mejores que los resultados de los modelo de árbol de decisión, este incremento favorable en los índices de performance no es tan significativo, mientras que, el tiempo de computo de los modelos de Random Forest si es notablemente mayor al resto.

Para encontrar los modelos óptimos de Random Forest también se utiliza la búsqueda en rejilla. Para ambas variables objetivo se observa en las Tablas 12 y 13 patrones similares. En primer lugar, para el

parámetro de “Ntree”, cantidad de árboles del bosque, se denota que a mayor cantidad de árboles se obtienen mejores indicadores, sin embargo, el tiempo de cómputo también es significativamente mayor. Se prueban modelos en el rango de [20,600] árboles. En cuánto al parámetro de “Mtry” se observa que utilizando la raíz cuadrada de la cantidad de variables de la base se obtienen los mejores resultados. Para el caso del parámetro “NodeSize” se explora entre valores de rangos de [20,100] para ambas variables objetivo. Por último se observa que cuando se utilizan técnicas de submuestreo, el parámetro de “SampleSize”, el rendimiento de los modelos es menos favorable.

Para el caso de la variable objetivo “Desempeño en Lengua”, se presenta en la Tabla 12, que el modelo con índices más asertivos es el número 10, que utiliza un valor de “Ntree” de 600, un “Mtry” de 13 y un “NodeSize” de 30 registros. Este modelo presenta una asertividad del 0.7371, con una tasa de error “OBB” del 0.2623. Sin embargo el tiempo de cómputo de este modelo es significativamente mayor a los modelos de árbol de decisión o XGBOOST.

En la Tabla 13 se exponen los resultados de los modelos de Random Forest para la variable objetivo de Matemática. El modelo de mejor rendimiento es el número 7, que trabaja con 500 árboles de decisión, un valor de “Mtry” de 13 y un valor de 50 como registros mínimos que tiene que haber en un nodo para realizar una división. Este presenta una exactitud de 0.7267 y una tasa de error “OBB” de 0.2948.

Tabla 12. Parámetros utilizados y resultados de los modelos la variable objetivo Lengua:

Nro. Modelo	Ntree	Mtry	Node_size	Sample Size	Max Nodes	ACC.	Recall	Precisión	Tasa de error OBB	Variable Objetivo
1	20	10	50	-	.	0.7289	0.5654	0.8319	28.38%	Lengua
2	50	10	50	-	-	0.7337	0.5608	0.8427	27.17%	Lengua
3	50	10	20	-	-	0.733	0.5603	0.8417	27.66%	Lengua
4	50	13	30	-	-	0.7338	0.5687	0.8377	27.45%	Lengua
5	200	13	30	-	-	0.7374	0.5660	0.8422	26.51%	Lengua
6	200	13	30	300000	5000	0.7161	0.5558	0.8486	28.31%	Lengua
7	200	13	30	300000	10000	0.7113	0.4902	0.8505	28.78%	Lengua
8	500	13	-	-	-	0.7366	0.5666	0.8437	26.33%	Lengua
9	500	13	30			0.7368	0.5834	0.8334	26.28%	Lengua
10	600	13	30	-	-	0.7371	58.40	0.8334	26.23%	Lengua

Tabla 13. Parámetros utilizados y resultados de los modelos la variable objetivo de Matemática:

Nro. Modelo	Ntree	Mtry	Node_size	Sample Size	Max Nodes	ACC.	Recall	Precisión	Tasa de error OBB	Variable Objetivo
1	10	13	30	-	-	0.6895	0.5905	0.7579	33.83%	Mat.
2	50	13	30	-	-	0.7105	0.5924	0.7920	29.76%	Mat.
3	50	10	30	-	-	0.7093	0.5848	0.7959	29.76%	Mat.
4	50	13	50	-	-	0.7119	0.5843	0.7999	29.53%	Mat.
5	50	13	100	-	-	0.7103	0.5689	0.8078	29.21%	Mat.
6	100	13	50	-	-	0.7136	0.5854	0.8020	28.92%	Mat.
7	500	13	30	-	-	0.7267	0.5978	0.8202	29.48%	Mat.

Resultados de los modelos óptimos de Random Forest utilizando los conjuntos de entrenamiento:

Si bien los modelos al predecir con los conjuntos de entrenamiento, presentan índices más favorables, estos no denotan una diferencia tal que con aquellos índices calculados con las predicciones del conjunto de testeo. No se observa un sobre ajuste de los modelos a los datos con los cuáles aprendieron inicialmente.

Tabla 14. Resultados de los modelos óptimos utilizando los conjuntos de entrenamiento:

					Conjunto de Testeo			Conjunto de Entrenamiento		
Variable Objetivo	**Nro. Modelo**	**Ntree**	**Mtry**	**Node_size**	**Accuracy**	**Recall**	**Precisión**	**Accuracy**	**Recall**	**Precisión**
Lengua	10	600	13	30	0.7371	0.5840	0.8334	0.8079	0.6687	0.9003
Mat.	7	500	13	30	0.7267	0.5978	0.8202	0.7962	0.6865	0.8912

iii. Parámetros y resultados tercer modelo: Extreme Gradient Boosting.

Cómo se observa en las Tabla 15 y 16, los resultados de los modelos de XGBoost, tanto para la variable Lengua como Matemáticas, presentan índices de performance con un leve incremento favorable respecto de los modelos anteriores. Esto es una secuencia lógica dado que este algoritmo es una versión mejorada de los anteriores. Estos modelos, no solo tienen un rendimiento más asertivo, si no que su tiempo de cómputo es significativamente menor que los otros.

Para la variable objetivo de Lengua, se explora entre el rango [50,500] para el parámetro "NRounds". En cuanto a la tasa de aprendizaje o "ETA" se prueba con valores de 0.1, 0.2 y 0.3. Para el caso de la profundidad de los árboles con los que trabajan los modelos se realiza una búsqueda entre el rango [6,10]. Asimismo se explora utilizando el 80% y 90% de columnas o registros para realizar los modelos y analizar cómo estas técnicas de submuestreo impactan en los resultados.

En cuánto a la variable de matemática, se explora entre el rango [125, 1500] para indicar la cantidad de secuencias que debe realizar el modelo. También se explora con valores de 0.1, 0.2 y 0.3 para la tasa de aprendizaje. Se realiza una búsqueda de rejilla entre el rango [6,10] para el parámetro de "MaxDepth" y por último, también se explora con técnicas de submuestreo.

En la Tabla 15 destaca el ejemplo número 11 para la variable de Lengua. Este presenta una precisión del 0.7564 con una tasa de Recall de

0.6292 y un índice de precisión de 0.8369. Este modelo utiliza una Tasa de aprendizaje del 0.3, un valor de "Max_depth" de 6, para la construcción de los diferentes árboles se utilizaron todas las variables y registros. Asimismo, se observa que el valor óptimo del parámetro "Nrounds" es 125, valores mayores en este parámetro no aumentan la eficiencia del mismo, mientras que si aumentan su tiempo de cómputo.

Por otro lado, para la variable objetivo de matemática, el modelo predilecto es el número 6. El mismo trabaja con una tasa de aprendizaje del 0.2, un valor de "Max_depth" de 6 y un valor de sub-muestro de columnas de 0.9. El mismo itera 125 veces. Este modelo presenta un "Accuracy " de 0.7564, con una sensibilidad de 0.6292 y una especificidad de 0.8369.

Tabla 15. Parámetros utilizados y resultados de los modelos la variable objetivo de Lengua:

Nro. Modelo	N Rounds	ETA	Max Depth	Sub Sample	Col Sample by tree	Min Child Weight	ACCURACY	Recall	Precisión	Variable Objetivo
1	100	0.3	6	-	-	-	0.7458	0.5892	0.8449	Lengua
2	100	0.3	-	-	0.7858	10	0.7463	0.5982	0.84	Lengua
3	100	0.3	6	1	-	1	0.7466	0.6	0.8392	Lengua
4	100	0.1	6	0.9	1	1	0.752	0.6047	0.8451	Lengua
5	100	0.3	6	0.9	1	1	0.7541	0.6245	0.8361	Lengua
6	100	0.3	6	1	1	1	0.7561	0.6259	0.8386	Lengua
7	200	0.3	6	1	1	1	0.7552	0.6296	0.8346	Lengua
7	300	0.3	6	1	1	1	0.7532	0.6274	0.8328	Lengua
8	500	0.3	6	0.8	1	0.8	0.7473	0.6219	0.8266	Lengua
9	50	0.3	6	1	1	1	0.754	0.6164	0.8411	Lengua
10	150	0.3	6	1	1	1	0.756	0.629	0.8363	Lengua
11	125	0.3	6	1	1	1	0.7564	0.6292	0.8369	Lengua
12	125	0.3	10	1	1	1	0.7422	0.6191	0.8201	Lengua
13	125	0.3	8	1	1	1	0.7488	0.6214	0.8295	Lengua
14	125	0.3	7	1	1	1	0.7525	0.6246	0.8334	Lengua

Tabla 16. Parámetros utilizados y resultados de los modelos la variable objetivo de Matemática:

Nro. Modelo	N Rounds	ETA	Max Depth	Sub Sample	Col Sample by tree	Min Child Weight	ACCURACY	Recall	Precisión	Variable Objetivo
1	125	0.3	6	1	1	1	0.728	0.6119	0.8081	Matemática
2	200	0.3	6	1	1	1	0.7269	0.6123	0.8060	Matemática
3	125	0.3	10	1	1	1	0.7151	0.6076	0.7894	Matemática
4	125	0.1	6	1	1	1	0.7275	0.6	0.8152	Matemática
5	125	0.2	6	1	1	1	0.73	0.6114	0.8119	Matemática
6	125	0.2	6	1	0.9	1	0.7308	0.6110	0.8135	Matemática
7	125	0.2	6	1	0.9	1	0.7297	0.6103	0.8121	Matemática
8	125	0.2	6	0.9	0.9	1	0.7291	0.6107	0.8110	Matemática
9	125	0.2	6	1	0.9	0.9	0.7292	0.6106	0.8112	Matemática
10	125	0.2	5	1	1	0.9	0.7297	0.6	0.8130	Matemática
11	300	0.2	6	1	0.9	1	0.7298	0.6164	0.8080	Matemática
12	1500	0.2	6	1	0.9	1	0.7235	0.6159	0.7982	Matemática

Resultados de los modelos óptimos de XGBOOST utilizando los conjuntos de entrenamiento:

Se observa en la Tabla 17, igual que con los modelos anteriores, los resultados de los modelos óptimos utilizando los conjuntos de entrenamiento no presentan índices significativamente más asertivos que cuando se los utilizan con los conjuntos de testeo. Se comprueba que no hay Overfitting en los modelos óptimos.

Tabla 17. Resultados de los modelos óptimos utilizando los conjuntos de entrenamiento:

								Conjunto de Testeo			Conjunto de Entrenamiento		
Var. Obj.	Nro	N Rounds	ETA	Max Depth	Sub Sample	Col Sample	Min Child Weight	ACC.	Recall	Precisión	ACC.	Recall	Precisión
L	11	125	0.3	6	1	1	1	0.7564	0.6292	0.8369	0.7855	0.6634	0.8622
M	6	125	0.2	6	1	0.9	1	0.7308	0.6110	0.8135	0.7528	0.6379	0.8321

e. Variables de mayor importancia en los modelos óptimos:

En las tablas 18 y 19 se destacan aquellas variables que presentan mayor preponderancia en los modelos óptimos de Lengua y Matemática. Para cada variable objetivo se seleccionan los ejemplares con mejores índices para analizar aquellas variables que tienen mayor peso en el rendimiento de los alumnos. Estos modelos óptimos fueron aquellos que utilizan el algoritmo de Extreme Gradient Boosting.

En las Tablas 18 y 19 se puede observar que ambas variables objetivo presentan como variables independientes de valor aquellas referidas a cómo se siente el alumno en su escuela, si se siente que es incluido o se siente excluido. La relación del mismo tanto con sus docentes cómo con sus compañeros, es decir que el mismo se sienta miembro de una comunidad, afecta su rendimiento en la escuela.

Además, son importantes aquellas variables que hacen referencia a los elementos con los que cuenta el alumno para poder estudiar, cómo lo es una computadora. Es importante destacar que la base de datos con la que se realiza el presente trabajo es del año 2021, segundo año de la pandemia de COVID-19, donde un gran porcentaje de las clases fueron virtuales. Además, es importante señalar que para ambos modelos es importante la variable de si el alumno tiene cómo responsabilidad un trabajo a la par del estudio sea dentro de su seno familiar o por fuera de este.

En cuanto al modelo específico de predicción de la variable objetivo Lengua, se observa en la Tabla 18 que son importantes las variables que refieren a la percepción del alumno de su capacidad en esta materia. Asimismo, se destaca la variable del sector del alumno, que refiere a si el estudiante va a una escuela privada o pública.

Por otro lado, para el caso de la variable objetivo Matemáticas, tienen mayor preponderancia la percepción del estudiante sobre su capacidad analítica y la provincia donde habita.

Tabla 18. Variables de mayor importancia modelo de Lengua:

Variables para el modelo de predicción "Desempeño en Lengua"		
Ranking	Variable	Descripción
1	ap34d	Si el alumno se siente incómodo o fuera de lugar en la escuela.
2	ap34c	Si el alumno siente que lo excluyen en la escuela.
3	ap20a	Si el alumno además de asistir a la escuela, trabaja con la familia.
4	sector	Si el alumno asiste a una escuela privada o estatal.
5	ap34e	Si el alumno se siente solo en la escuela.
6	ap10L	Si en la casa del alumno hay servicios de conexión a Internet.
7	ap29a	Si el alumno no pudo asistir a clases virtuales durante la pandemia porque no tenía una computadora en su casa.
8	ap42	Si el alumno piensa que le va bien en su clase de Lengua.
9	ap43	Opinión del alumno sobre su lectura.
10	ap11d	Cuántas computadoras hay en la casa del alumno.

Tabla 19. Variables de mayor importancia modelo de Matemáticas:

Variables para el modelo de predicción "Desempeño en Matemática"		
Ranking	Variable	Descripción
1	ap34d	Si el alumno se siente incómodo o fuera de lugar en la escuela.
2	ap34e	Si el alumno se siente solo en la escuela.
3	ap45	Si el alumno piensa que le va bien en su clase de Matemáticas.
4	ap46	Opinión del alumno de su capacidad de resolver problemas matemáticos
5	ap34c	Si el alumno siente que lo excluyen en la escuela.
6	ap29a	Si el alumno no pudo asistir a clases virtuales durante la pandemia porque no tenía una computadora en su casa.
7	jurisdicción	Provincia del alumno.
8	ap37e	Si el alumno en el último mes jugó con videojuegos.
9	ap30	Cantidad de inasistencias en lo que va del año.
10	ap20b	Si el alumno además de ir a la escuela tiene algún trabajo por fuera de su entorno familiar

11. CONCLUSIONES FINALES.

El presente trabajo tiene como objetivo primario realizar un análisis original sobre la base de datos pública de las pruebas Aprender que realiza el Ministerio de Educación de la Nación. El mismo es de valor ya que detecta de forma clara aquellos factores específicos que afectan al rendimiento de los alumnos en las áreas de Lengua y Matemática.

Si bien en el campo de las áreas sociales es habitual realizar estudios en los cuales se buscan las causas a diferentes fenómenos de las esferas humanas, el presente trabajo tiene como novedad que se utilizan técnicas de la Ciencia de datos para la búsqueda de estas. Cómo se mencionó previamente, el uso de este tipo de técnicas y herramientas en las materias sociales no está ampliamente difundido, incluso menos en estudios sobre educación. Es de crucial importancia que cada vez más profesionales de diferentes sectores puedan beneficiarse de las ventajas de estas nuevas tecnologías.

Asimismo, es importante realizar un breve comentario sobre los factores destacados que afectan el desempeño de los alumnos. Los mismos se los podría clasificar en 3 grandes categorías para ambas variables.

La primera de estas, es cómo se sienten los estudiantes en sus escuelas, si se sienten parte de una comunidad o se sienten excluidos. La segunda categoría, es si los estudiantes poseen los elementos materiales necesarios para llevar adelante sus estudios. Esto es lógico dado que las pruebas se realizaron en el año 2021, segundo año de pandemia donde la mayor parte del año las clases fueron virtuales. Por último, la tercera categoría, es si los alumnos tienen un trabajo cómo responsabilidad a par de sus estudios. Tener esta información es importante dado que aquellas personas con la potestad para hacerlo pueden trabajar sobre estas áreas de influencia para mejorar los niveles de educación de las futuras generaciones.

Por último, este trabajo utiliza tres algoritmos de la familia de árboles de decisión. Cómo se expone en este trabajo, los modelos óptimos para trabajar con la base de datos de los resultados de las pruebas Aprender son los modelos de Extreme Gradient Boosting, no solo por los resultados de performance de los mismos, si no también por su rapidez de ejecución.

12. REFERENCIAS.

Alaminos-Fernández, A.F. (2022). Árboles de decisión en R con Random Forest. Obets Ciencia Abierta. Limencop S. L. ISBN: 978-84-09-49460-6, Págs: 36-40. http://hdl.handle.net/10045/133067

Breiman, L. (2001). Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Centeno, C. P., & Leal, M. (2011). ¿Han funcionado las reformas educativas en América Latina? Un estudio de los casos de Argentina, Brasil y Chile. Archivos Analíticos de Política Educativa, 19(0), 36. https://doi.org/10.14507/epaa.v19n36.2011

Chan, D., Gabriela Galli, M., & Galli, D. (2020). Aplicación de técnicas estadísticas multivariadas con el lenguaje de programación R en investigaciones educativas del nivel superior. RAES: Revista Argentina de Educación Superior, ISSN-e 1852-8171, No. 20, 2020, Págs. 123-136, 20, 123–136. https://dialnet.unirioja.es/servlet/articulo?codigo=7592065&info=resumen&idioma=ENG

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd international conference on knowledge discovery and data mining (pp. 785-794).

Comisión Económica para América Latina y el Caribe. (2020). La educación en tiempos de la pandemia de COVID-19. Retrieved April 11, 2023, from https://www.siteal.iiep.unesco.org/respuestas_educativas_covid_19.

Dirección de Información Educativa. (2021). Cuadros indicadores de proceso 2020-2021 según provincias. [Archivo de datos]. Retrieved March 23, 2023, from https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.argentina.gob.ar%2Fsites%2Fdefault%2Ffiles%2F2018%2F04%2Fcuadros_indicadores_de_proceso_2020-2021_segun_estru_prov_041122.xlsx&wdOrigin=BROWSELINK

Engel, U., Quan-Haase, A., Xun Liu, S., Lyberg, L. (2022). Handbook of Computational Social Science, Volume 1. DOI: 10.4324/9781003024583-10

Giuffré, L., & Ratto, S. E. (2013). Applicable Quality Models in Higher Education in Argentina. Scientific Research, 4(10A), 29–32. https://doi.org/10.4236/ce.2013.410A005

Haro Rivera M. (2020). Árbol De Decisión, Aplicación Con Datos Meteorológicos/Decision Tree, Application With Meteorological Data. KnE Engineering, 5(2), 37–46. https://doi.org/10.18502/keg.v5i2.621

Honorable Congreso de la Nación Argentina. (2006). Ley de Educación Nacional. Retrieved March 18, 2023, from https://www.argentina.gob.ar/normativa/nacional/ley-26206-123542/actualizacion

IBM Corp. (2020). IBM SPSS Statistics for Windows (Version 27.0) [Software] IBM Corp. https://www.ibm.com/products/spss-statistics

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news ISSN 1609-3631, 2(3), 18-22.

Ministerio de educación de la nación. (2022). Informe Nacional de resultados: Análisis sobre los logros de aprendizaje y sus condiciones. Retrieved April 11, 2023, from https://www.argentina.gob.ar/sites/default/files/informe_aprender_2021_1.pdf

Naciones Unidas en Argentina. (2021). Análisis Conjunto del Sistema de Naciones Unidas 2021: Los efectos de la pandemia por Covid-19 en la Argentina. Retrieved April 11, 2023, from https://argentina.un.org/es/145708-an%C3%A1lisis-conjunto-del-sistema-de-naciones-unidas-2021-los-efectos-de-la-pandemia-por-covid

Narodowski, M., Moschetti, M., Gottau, V. (2017). El crecimiento de la educación privada en Argentina: Ocho explicaciones paradigmáticas. Cadernos de Pesquisa. https://doi.org/10.1590/198053143853

Piovani, J. I. (2022). Las desigualdades educativas en Argentina: análisis sincrónico de la situación y trayectoria escolar de diferentes cohortes de niños y adolescentes. Civitas - Revista de Ciências Sociais, 22, e41864. https://doi.org/10.15448/1984-7289.2022.1.41864

Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. International Journal of current engineering and technology, 3(2), 334-337.

R Core Team (2023) R: A language and environment for statistical computing (Version 4.3.1) [Software] R Foundation for Statistical Computing. https://www.r-project.org/

SAS Institute Inc. (2023) SAS VIYA Statistical Analysis Software. (Version 9.4M8) [Software] SAS Institute Corp. https://www.sas.com/en_us/software/viya.html

STATA Corp. (2023). STATA Statistical Software for Data Science (Version 18.0) [Software] Stata Corp. https://www.stata.com/

Zengin, K., Esgi, N., Erginer, E., Aksoy, M. E. (2011). A sample study on applying data mining research techniques in educational science: developing a more meaning of data. Procedia Social and Behavioral Sciences 15 (2011) 4028–4032. doi:10.1016/j.sbspro