PERÚ Ministerio de Educación

OSEE
OFICINA DE SEGUIMIENTO Y EVALUACIÓN ESTRATÉGICA

INFORME

Metodología de proyección de la matrícula escolar, secciones y docentes en el Perú: Una aproximación con Machine Learning

Febrero 2024

METODOLOGÍA DE PROYECCIÓN DE LA MATRÍCULA ESCOLAR, SECCIONES Y DOCENTES EN EL PERÚ

Una aproximación con *Machine Learning*

Jefa de la Oficina de Seguimiento y Evaluación Estratégica

Lourdes Patricia Vargas Vilchez.

Elaboración de contenidos:

- Pedro Emilse Casaverde Ayma
- Cristian Dominico Centeno Guzmán
- Erik Carl Candela Rojas

MINISTERIO DE EDUCACIÓN

Oficina de Seguimiento y Evaluación Estratégica

Febrero - 2024

Sede Central: Calle Del Comercio N° 193 Lima - Lima - San Borja - 15021 Perú

Teléfono: (01) 615-5800

https://www.gob.pe/minedu

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE ILUSTRACIONES

RESUMEN

El presente informe detalla la metodología empleada para desarrollar los modelos analíticos basados en técnicas de Aprendizaje Automático (*Machine Learning*, ML), los cuales fueron ajustados para proyectar, con un horizonte temporal de un año, la «matrícula escolar, las secciones y docentes» para cada grado escolar de los servicios educativos de la «Educación Básica Regular» (EBR) en el Perú. La elaboración de los modelos se fundamentó principalmente de los datos administrativos proporcionados por el Ministerio de Educación (Minedu), los cuales revelaron su idoneidad para la implementación de modelos de ML.

Las proyecciones generadas por los modelos de ML fueron sometidas a una evaluación, comparándolas con las proyecciones realizadas por la Unidad de Estadística (UE) de la Oficina de Seguimiento y Evaluación Estratégica (OSEE) del Minedu para el año 2021. La métrica empleada para evaluar el rendimiento de los modelos fue la «Raíz Cuadrada del Error Cuadrático Medio» (*Root Mean Square Error*, RMSE) aplicada a ambas proyecciones. Como resultado de esta evaluación, se evidenció que los modelos de ML lograron reducir el RMSE, especialmente en los niveles de primaria y secundaria de EBR.

Palabras Claves: Proyección de Matrícula Escolar, Proyección de la Demanda Educativa, Proyección de Secciones, Proyección de Docentes, Planificación Escolar, Planificación Educativa, *Machine Learning* en Educación.

ABSTRACT

This report outlines the methodology used to develop analytical models employing Machine Learning (ML) techniques, finely tuned to forecast, within a one-year timeframe, the «enrollment, sections, and teachers» for each grade level in Peru's Regular Basic Education (EBR) services. The development of these models was primarily grounded in administrative data provided by the Ministry of Education (Minedu), showcasing its suitability for implementing ML models.

The projections produced by the ML models underwent an evaluation, comparing them with the projections from the Statistics Unit (UE) of the Office of Monitoring and Strategic Evaluation (OSEE) at Minedu for the year 2021. The metric used to assess model performance was the Root Mean Square Error (RMSE) applied to both sets of projections. The outcome of this evaluation revealed that the ML models effectively reduced RMSE, particularly in the primary and secondary levels of EBR.

Keywords: School Enrollment Projection, Educational Demand Projection, Section Projection, Teacher Projection, School Planning, Educational Planning, Machine Learning in Education.

INTRODUCCIÓN

La proyección de la «matrícula escolar»[1] desempeña un papel fundamental en la «planificación educativa»[2] porque sirve como una herramienta para ayuda a anticipar las futuras necesidades que podrían requerir las instituciones educativas. Por ende, contar con estas proyecciones mejora el proceso de asignación de recursos en las instituciones educativas, propiciando una planificación más eficiente de la oferta educativa.

En el Perú, la proyección de la «matrícula escolar, las secciones y los docentes» (demanda educativa) en cada «código modular»[3] y grado de las instituciones Educativas está a cargo de la «Unidad de Estadística» (UE) de la «Oficina de Seguimiento y Evaluación Estratégica» (OSEE) del «Ministerio de Educación del Perú» (Minedu). Esta proyección emplea principalmente los datos obtenidos del «Censo Educativo»[4] y del «Sistema de Información de Apoyo a la Gestión Educativa» (SIAGIE).

Cabe destacar que lograr proyecciones precisas se convierte en un desafío considerable en un país caracterizado por la elevada movilidad estudiantil, cambios demográficos constantes y la presencia de factores económicos imprevisibles, entre otros. Estas particularidades incrementan el nivel de incertidumbre en las proyecciones, resaltando así la relevancia de perfeccionar constantemente las metodologías de pronóstico.

En este contexto, surge como una alternativa innovadora el desarrollo de un nuevo modelo de *Machine Learning* (ML) destinado a mejorar la metodología de pronóstico. La razón principal detrás de la adopción de esta nueva aproximación radica en la capacidad del ML para modelar de manera más precisa y dinámica las múltiples variables que influyen en la demanda educativa, superando las limitaciones inherentes de los enfoques tradicionales.

El objetivo principal de este informe es presentar la nueva metodología utilizada para proyectar la demanda educativa de la EBR en el Perú, empleando técnicas basadas en algoritmos de ML para estimar el nuevo modelo de proyección. Además, se detallarán las etapas de desarrollo y las métricas de robustez obtenidas de la nueva metodología.

---

[1] Conjunto de etapas que tienen por finalidad realizar la inscripción de un estudiante en una institución educativa o programa (Minedu, 2022a).
[2] Actividad que permite a las autoridades públicas guiar el desarrollo educativo y determinar las intervenciones prioritarias (Wright, 2015)
[3] Código que identifica al servicio educativo.
[4] Proceso estadístico que se realiza cada año en la UE, donde se recoge información detallada de las Instituciones Educativas (públicas y privadas) y Programas No Escolarizados en el Perú.

El modelo se creó siguiendo el marco de trabajo «Cross-Industry Standard Process for Data Mining» (CRISP-DM), una metodología ampliamente reconocida en proyectos vinculados con minería de datos, ciencia de datos e inteligencia artificial. La estructura de este informe se basa en las etapas que abarca la metodología CRISP-DM, desde la comprensión del problema hasta la evaluación de los resultados.

El primer capítulo, denominado «Comprensión del Problema», aborda aspectos esenciales como el objetivo principal, la situación actual y una revisión exhaustiva de la literatura. En el segundo capítulo, «Comprensión de los Datos», se detallan las fuentes de información, el proceso de recopilación de datos y los análisis exploratorios realizados. Posteriormente, en el tercer capítulo, «Preparación de los Datos», se aborda la conversión de los datos analizados en información, a través de un proceso que incluye aspectos como el tratamiento de valores faltantes, manejo de valores múltiples y estandarización de datos, entre otros. El cuarto capítulo, «Modelamiento», cubre todas las áreas relacionadas con el desarrollo de modelos de *Machine Learning* para la proyección de la demanda educativa. Finalmente, en el quinto capítulo, bajo el título «Evaluación de Resultados», se explora la evaluación realizada mediante la métrica de desempeño RMSE, contrastándola con la métrica obtenida a través de la metodología actual de la UE para el año 2021. Las conclusiones y recomendaciones derivadas del informe se presentan al cierre de este análisis.

CAPÍTULO I: COMPRENSIÓN DEL PROBLEMA

1.1. DETERMINACIÓN DE OBJETIVOS QUE BUSCA LA GESTIÓN

En la fase inicial de la metodología CRISP-DM, resulta fundamental definir con absoluta precisión el objetivo que busca la «Unidad de Estadística» (UE) de la «Oficina de Seguimiento y Evaluación Estratégica» (OSEE) con respecto a esta nueva metodología de proyección. Este objetivo debe actuar como un punto de partida para fomentar la innovación y verificar la coherencia de los resultados obtenidos con dicho objetivo. En este contexto, se describe el objetivo de gestión de la siguiente manera:

Objetivo de Gestión: Mejorar la precisión en la proyección de la matrícula escolar, las secciones y los docentes de las instituciones educativas de Educación Básica Regular (EBR) en el Perú.

Justificación: Mejorar la precisión en las proyecciones es crucial para planificar mejor y asignar recursos de manera más efectiva en las instituciones educativas, asegurando así una respuesta adecuada a la demanda educativa futura.

Asimismo, la implementación de esta nueva metodología de pronóstico está alineado con la finalidad pública del Minedu, específicamente con el Objetivo Estratégico Institucional (OEI) y su correspondiente Acción Estratégica Institucional (AEI), los cuales se describen a continuación:

- **OEI 6:** Modernizar la gestión y financiamiento institucional y del sistema educativo.
- **AEI 6.6:** Estrategias efectivas para la implementación de políticas y toma de decisiones con las instancias vinculadas al sector educación.

1.2. EVALUACIÓN DE LA SITUACIÓN ACTUAL

Previo a la determinación de los objetivos más técnicos, es crucial tener en cuenta que la UE lleva a cabo la proyección de la matrícula escolar, las secciones y los docentes en todos los niveles de EBR, que incluyen el nivel inicial, primaria y secundaria. En este proceso, se utilizan tanto técnicas de regresión lineal como no lineal, con el objetivo de minimizar error cuadrático medio. A partir de esto, se generan un conjunto de nueve modelos donde se considera el «año» como parte de las variables predictoras (x_n), mientras que el «número de estudiantes matriculados» se convierte en la variable objetivo (y).

Tabla 1:

Modelos usados por la UE para la proyección de la matrícula escolar.

Modelo	Expresión
Modelo 1	$Y_{ijt} = \alpha + \beta X + \mu_{ijt}$
Modelo 2	$Y_{ijt} = \alpha + \beta \ln X + \mu_{ijt}$
Modelo 3	$Y_{ijt} = \alpha + \beta(1/X) + \mu_{ijt}$
Modelo 4	$\ln Yijt = \alpha + \beta X + \mu ij$
Modelo 5	$\ln Y_{ijt} = \alpha + \beta \ln X + \mu_{ijt}$
Modelo 6	$\ln Y_{ijt} = \alpha + \beta(1/X) + \mu_{ijt}$
Modelo 7	$1/Y_{ijt} = \alpha + \beta X + \mu_{ijt}$
Modelo 8	$1/Y_{ijt} = \alpha + \beta \ln X + \mu_{ijt}$
Modelo 9	$1/Yijt = \alpha + \beta(1/X) + \mu ijt$

Donde:

- Y_{ijt}: Matrícula del i-ésimo grado (edad) en el j-ésimo servicio educativo del año «t».
- X: Año «t» (2018 – 2024).
- μ_{ijt}: Errores de estimación de la matrícula del i-ésimo grado (edad) en el j-ésimo servicio educativo del año «t».

Para la generación de los modelos y la proyección se hace uso principalmente de las siguientes bases de datos (último proceso de proyección para los años 2023 y 2024):

- Datos del SIAGIE, para el período 2016[5] – 2022 (corte de datos al mes de abril).
- El padrón de servicios educativos actualizados al 19 de abril de 2022.
- Censo Educativo de los períodos 2016-2021 (datos de EBR).

El proceso de modelado y las proyecciones comprende las siguientes 2 etapas:

a. **Etapa I:** Se realiza la proyección de la matrícula escolar, las secciones y los docentes para todos los grados escolares por código modular. Cabe precisar que la proyección de las secciones sirve como una aproximación o sustituto para estimar la cantidad de docentes requeridos (Minedu, 2022b).
b. **Etapa II:** Se reajusta las proyecciones donde se comprueban que no haya proyecciones de códigos modulares y grados con cantidades de matrículas de alumnos iguales a cero. Además, se realizan los redondeos de las proyecciones a

[5] Es relevante mencionar que en el caso del SIAGIE, se han considerado años anteriores que han contribuido en la estimación.

cantidades enteras (dígito superior) y las proyecciones que son incoherentes (menores a las reales en el último año) son reajustadas con base en la matrícula del último año.

Es importante resaltar que, la metodología utilizada por la UE ha acumulado una experiencia de ejecución durante los últimos cinco años, por lo que evidencia su robustez a nivel metodológico. Sin embargo, emplear el «número de estudiantes matriculados» para ajustar una curva que pueda proyectar el crecimiento futuro de la matrícula no siempre es la premisa más adecuada, ya que podría haber patrones de crecimiento inusuales recientes en la población (Armstrong & Nunley, 1981).

Por esta razón, resulta esencial tener en cuenta múltiples variables predictoras que puedan estar vinculadas con la variable objetivo. Este enfoque, se denomina regresión lineal múltiple, donde se incorporan diversas variables explicativas para modelar la relación entre ellas y la variable objetivo. Al hacerlo, se considera una variedad más extensa de factores que pueden incidir en la proyección de la demanda educativa, lo cual tiene el potencial de mejorar significativamente la precisión del modelo.

1.3. DETERMINACIÓN DE LOS OBJETIVOS TÉCNICOS

Los objetivos técnicos se establecen para abordar los desafíos identificados durante la evaluación de la situación inicial y se alinean con el objetivo de gestión establecido. Con base a esto, se establece los siguientes objetivos técnicos.

Objetivos Específicos:

A. Desarrollar un modelo de *Machine Learning* que permita la proyección[6], con un horizonte temporal de un año, de la matrícula escolar, las secciones y los docentes para cada grado en los servicios educativos de la Educación Básica Regular en el Perú.
B. El modelo de *Machine Learning* debe incorporar variables adicionales al «número de estudiantes matriculados» que contribuyan a mejorar la precisión en la proyección de la demanda educativa.
C. Reducir la raíz del error cuadrático medio por debajo del nivel actual en el modelo utilizado por la Unidad Estadística.[7]

[6] En relación con el objetivo específico «A», la estimación del modelo se llevará a cabo utilizando datos históricos de matrícula escolar correspondientes a los años 2018, 2019 y 2020. Los detalles completos del proceso de modelamiento se presentan en el capítulo IV.
[7] La validación se llevará a cabo analizando el error obtenido en la proyección para el año 2021.

1.4. REVISIÓN DE LA LITERATURA

Existen diversas investigaciones que han analizado la proyección de la demanda educativa en instituciones educativas públicas y privadas. A continuación, se detallan las investigaciones revisadas siguiendo un orden cronológico.

Desde las primeras investigaciones, se reconocía la importancia de anticipar la matrícula escolar para lograr un equilibrio eficaz entre la planificación de instalaciones educativas y la oferta y demanda de docentes. Un estudio realizado por la división de investigación de la «Asociación Nacional de Educación» (NEA, por sus siglas en inglés) señaló que, para proyectar la cantidad de estudiantes, se pueden emplear indicadores históricos, como la relación entre los nacimientos y la matrícula, o la proporción de la población en edad escolar con respecto a la matrícula. La investigación reveló que las proyecciones para comunidades con un crecimiento poblacional estable son más precisas que aquellas correspondientes a comunidades en rápido crecimiento. Además, se observó que las proyecciones de la matrícula escolar total tienden a ser más exactas que las proyecciones por grado o nivel educativo. Por último, el estudio insta a las autoridades escolares a solicitar y mantener actualizadas las proyecciones de matrícula, resaltando que la precisión de los resultados se verá mejorada a medida que se perfeccionen los datos fuente (Research Division, National Education Association, 1953, págs. 46-52).

En esa misma línea, la NEA señalaba que la proyección de la matrícula escolar permite estimar la cantidad de docentes que serán necesarios para educar las futuras generaciones. En su investigación, lograron identificar los siguientes procedimientos para estimar la cantidad de docentes por aula.

- **Opción 1:** Dividir la matrícula proyectada con el número promedio de alumnos matriculados por maestro.
- **Opción 2:** Dividir la matrícula proyectada con el número promedio de alumnos que asisten diariamente a sus clases por maestro.

Entre ambas opciones, la Opción 2 demostró proporcionar proyecciones más estables. No obstante, la demanda de maestros de aula no es homogénea en cada grado, y muchos docentes imparten clases en múltiples niveles, lo que hace que estos procedimientos resulten poco prácticos (Research Division, Nacional Education Association, 1953, págs. 53-58).

Con el transcurso del tiempo, Armstrong y Nunley (1981) llevaron a cabo un estudio en el cual se diseñaron dos métodos de proyección de matrícula para el colegio comunitario «*Montgomery College*», situado en el condado de *Montgomery, Maryland*. Estos métodos se configuraron de la siguiente manera:

- **Método I - Ajuste de curvas (Series de Tiempo):** Señala que el crecimiento de matrícula en un periodo de tiempo determinado puede describirse mediante algún tipo de función matemática, donde el crecimiento futuro de la matrícula continuara con la trayectoria que ha seguido en el pasado. Para este método se evaluaron tres tipos de curvas (una lineal y dos no lineales) para conocer que tan eficaces son para describir el crecimiento de la matrícula en *Montgomery College*.
- **Método II: Rendimiento desde componentes de población (Regresión):** En este método, los estudiantes de *Montgomery College* se dividen en tres grupos principales o componentes. El primer grupo, está compuesto por nuevos ingresantes, que son recientes graduados de secundaria del condado de *Montgomery* y áreas circundantes en *Maryland* y *Washington*, *D.C*. El segundo grupo lo conforman los estudiantes matriculados que regresan de semestres anteriores. Por último, el tercer grupo está compuesto por adultos residentes del condado de *Montgomery*. Para cada uno de estos componentes se obtiene los indicadores de tamaño referencial de su población y la tendencia para calcular su respectiva tasa de rendimiento. Con esta información, se aplica las tasas de rendimiento al tamaño de las poblaciones que se espera tener en el futuro para cada componente. De esta forma, se estima la matrícula total de un año determinado a partir de la suma de los componentes proyectados.

Como resultado de la evaluación, los investigadores concluyeron que el enfoque basado en componentes es el método más fiable y fundamentado de los dos. Sin embargo, los investigadores señalan que no se debe descartar el método de ajuste de curva, ya que puede utilizarse como una herramienta de validación para las proyecciones generadas mediante el enfoque basado en componentes. En caso de que surja una discrepancia significativa entre las proyecciones obtenidas a través de estos dos métodos, esto puede servir como una señal de alerta, indicando la necesidad de un análisis más profundo para identificar las causas subyacentes del problema.

En una investigación más reciente llevada a cabo por Hussar y Bailey (2011), se proyectaron variables educativas hasta el año 2020. Estas incluyeron la matrícula, el número de graduados, la cantidad de maestros y los gastos de escuelas públicas y privadas de educación primaria y secundaria. Además, se analizaron la matrícula y los

títulos conferidos en instituciones de educación superior que otorgan grados. Se examinaron detalladamente los datos nacionales de matrícula y graduados de los últimos 15 años, así como la información estatal relacionada con la matrícula en escuelas públicas de educación primaria y secundaria. Las proyecciones se fundamentaron principalmente en los siguientes dos métodos:

- Suavización exponencial: Esta técnica se utilizó en las proyecciones tanto de la matrícula en educación primaria y secundaria como en los graduados de escuela secundaria. Además, desempeñó un papel crucial en las estimaciones relacionadas con los docentes en los niveles de educación primaria y secundaria, así como en las proyecciones de matrícula y títulos conferidos a nivel de educación superior.
- Regresión lineal múltiple: Esta técnica se implementó para proyectar la cantidad de docentes y los gastos en los niveles de educación primaria y secundaria, así como en las proyecciones de matrícula y títulos otorgados a nivel de educación superior.

Posteriormente, Zhuang y Gan (2017) desarrollaron un estudio donde proyectaron la matrícula escolar de las «Escuelas Públicas de Chicago» (CPS, por sus siglas en inglés). Para llevar a cabo esta tarea, utilizaron la extensa base de datos privada de CPS, la cual alberga 400 GB de información, complementándola con datos disponibles en el sitio web oficial de CPS. Asimismo, se obtuvieron las características escolares, incluida la calificación de las escuelas, a través del portal de datos de Chicago. A partir de esta información, se desarrollaron 10 indicadores, que son:

- Distancia desde la dirección de casa hasta la escuela secundaria.
- Porcentaje de estudiantes del mismo grupo étnico.
- Porcentaje de estudiantes que necesitan inglés como segunda lengua.
- Porcentaje del mismo género.
- Calificación de escuela.
- Cantidad de estudiantes que entran y salen de una escuela.
- Tasa de asistencia promedio de una escuela.
- Dicotómica si la escuela es una de captación para estudiantes.
- Dicotómica: concordancia entre escuelas secundaria e intermedia del estudiante.
- Número de matrícula del año pasado.

Con estos indicadores se elaboró el modelo «regresión logística condicional», diseñado para mejorar la proyección de inscripciones en nuevas escuelas secundarias. Sin embargo, este modelo no funciona bien en las escuelas antiguas, por lo que se optó por replicar las inscripciones del año anterior debido a su alta correlación. Los resultados obtenidos superaron las proyecciones actuales de la CPS y de la regresión lineal.

Finalmente, en un estudio llevado a cabo por Reichardt et al. (2020), en colaboración con el Departamento de Educación Primaria y Secundaria de Missouri y el Laboratorio Regional de Educación Central, se desarrolló un modelo predictivo destinado a estimar el número de maestros que no están certificados de manera adecuada. Este indicador se utiliza como una métrica de la escasez de maestros. El modelo se basa en información actualizada sobre la matrícula escolar y en el número de maestros asignados en períodos académicos anteriores. Entre los datos históricos y actuales se encuentran el total de maestros empleados, los que abandonan sus distritos, las nuevas contrataciones y aquellos maestros que carecen de la certificación adecuada. En cuanto a los resultados derivados del modelo, se obtienen cifras referentes al número de maestros empleados, los que abandonan, las nuevas contrataciones y los maestros que no cuentan con la certificación requerida.

Finalmente, una variable directamente relacionada con la matrícula escolar es la deserción escolar interanual, que indica que un estudiante matriculado en el año T no se matriculara en el año T+1, reduciendo así la tasa de matrícula escolar. Candela y Centeno (2022) llevaron a cabo la estimación de modelos de ML para calcular el riesgo de interrupción de estudios que tienen los estudiantes de los servicios educativos de EBR en el Perú. Los resultados revelaron que las variables que contribuyen a la predictibilidad de la interrupción de estudios varían según el nivel educativo. En el nivel inicial, se observó que las variables relacionadas con la extra-edad, la participación en programas de transferencia monetaria condicionada, como el Programa JUNTOS, y el rendimiento académico en matemáticas son algunas de las más influyentes en la predictibilidad. En el nivel primaria, destacaron variables como la participación en el programa JUNTOS, la extra-edad y la proyección de ingresos del hogar. Por último, en el nivel de secundaria, las variables que más contribuyeron fueron el rendimiento académico en comunicación y matemáticas, y la proyección de ingresos en el hogar del estudiante.

CAPÍTULO II: COMPRENSIÓN DE LOS DATOS

2.1. RECOPILACIÓN DE DATOS INICIALES

En el punto 1.4, «REVISIÓN DE LA LITERATURA», se pudo identificar diversas variables relacionadas con la proyección de la matrícula escolar, las secciones y docentes, las cuales se listan a continuación:

- **Research Division, National Education Association (1953, págs. 46-52):** Tamaño de población, matrícula escolar y nacimientos.
- **Research Division, Nacional Education Association (1953, págs. 53-58):** Asistencia a clases y matrícula escolar.
- **Armstrong y Nunley (1981):** Matrícula escolar y tamaño de población.
- **Hussar y Bailey (2011):** Matrícula escolar, docentes.
- **Zhuang y Gan (2017):** Demografía estudiantil, asistencia escolar, accesibilidad al servicio educativo, rendimiento académico y matrícula escolar.
- **Reichardt et al. (2020):** Docentes, situación laboral de docentes, matrícula escolar y migración.
- **Candela y Centeno (2022):** Demografía estudiantil, situación económica del estudiante y rendimiento académico.

Todas estas variables guardan una relación directa con los datos de los estudiantes, los servicios educativos, la población y los docentes. En este contexto, se llegó a la conclusión de que la obtención de información sobre los estudiantes podía realizarse mediante el Censo Educativo y el «Sistema de Información de Apoyo a la Gestión de la Institución Educativa» (SIAGIE). Respecto a los datos relativos a los servicios educativos, se determinó que era posible consultar el «Padrón de Instituciones Educativas» y el «Censo Educativo». En lo que respecta a la información sobre la población, se identificó la viabilidad de obtenerla a través del Censo Nacional del «Instituto Nacional de Estadística e Informática» (INEI), el «Sistema Informático Nacional de Defunciones» (SINADEF) y los datos de Población Estimada por Edades del «Ministerio de Salud del Perú» (Minsa).

En cuanto a los datos de los docentes, se decidió no considerarlos directamente, ya que se optó por emplear el método de proyección de secciones para estimar la proyección de docentes. Cabe destacar que esta decisión se alinea con el criterio utilizado por la UE para estimar los docentes requeridos (Minedu, 2022b). La Tabla 2 detalla las fuentes de información mencionadas.

Tabla 2:

Fuentes de información.

N	Descripción
1	**Censo Educativo:** Estudiantes y secciones de los años 2018, 2019 y 2020. *Link:* http://escale.minedu.gob.pe/censo-escolar-eol/
2	**SIAGIE:** Estudiantes y secciones de los años 2018, 2019 y 2020. *Link:* https://siagie.minedu.gob.pe/inicio/
3	**Padrón Instituciones Educativas:** Servicios educativos de EBR de los años 2018, 2019 y 2020. *Link:* http://escale.minedu.gob.pe/uee/-/document_library_display/GMv7/view/958881
4	**Población Estimada por Edades:** Población estimada por edades simples a nivel distrital del año 2017 al 2020. *Link: https://www.minsa.gob.pe/reunis/data/poblacion_estimada.asp*
5	**SINADEF:** Fallecidos del año 2017 al 2020. *Link: https://www.minsa.gob.pe/defunciones/?op=1*
6	**Censo Nacional del INEI:** Analfabetismo, población proyectada, ausentismo y población adulta con secundaria completa a nivel distrital del año 2017. *Link: http://censo2017.inei.gob.pe/resultados-definitivos-de-los-censos-nacionales-2017*

2.2. DESCRIPCIÓN DE LOS DATOS

A partir de las fuentes de información recopiladas, se logró identificar variables que están en consonancia con la literatura revisada. Estas variables, una vez identificadas, fueron organizadas en tres categorías, tal como se muestra en la tabla 3.

Tabla 3:

Grupos de variables.

Grupo	Variables[8]
Estudiantes	Matrícula escolar.
Servicios educativos	Secciones, grados, distancias con relación a la ubicación del servicio educativo.
Población	Población a nivel distrital, ausentismo, analfabetismo, secundaria completa, fallecidos.

[8] La Tabla 4 presenta de manera detallada la descripción de las variables identificadas.

2.3. EXPLORACIÓN DE LOS DATOS

2.3.1. SERIES DE MATRÍCULA ESCOLAR POR NIVEL EDUCATIVO

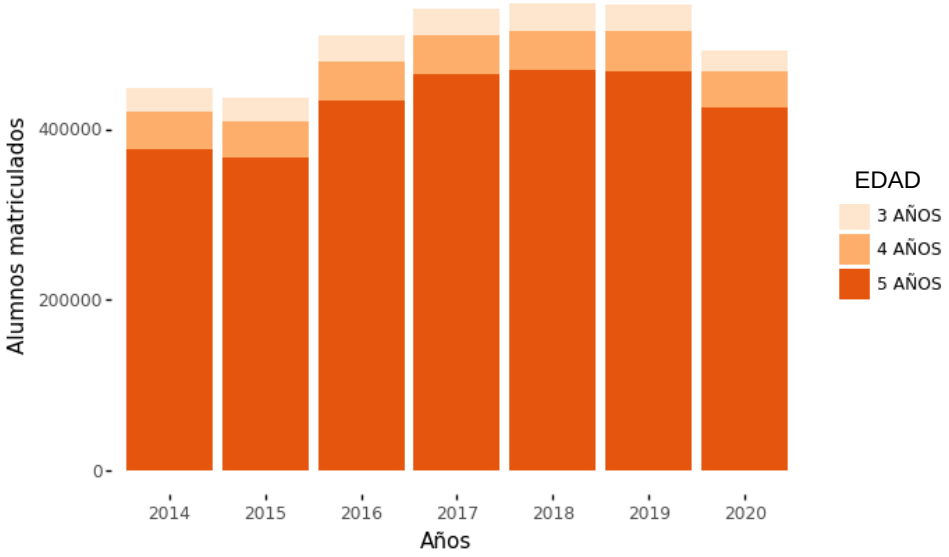
En esta etapa, se examinan los datos en relación con la variable de análisis, que en este caso es la cantidad de matrículas por grado y nivel educativo de EBR.

Para lograrlo, es necesario examinar los datos históricos de los años 2014, 2015, 2016, 2017, 2018, 2019 y 2020, relacionados con las matrículas por nivel educativo y grado escolar correspondiente.

En el Gráfico 1, se aprecia que, para el nivel Inicial, la cantidad de alumnos matriculados experimenta un crecimiento sostenido desde 2014 hasta 2017. Entre 2018 y 2019, parece mantenerse en una tendencia estable. Sin embargo, en el año 2020, se registra una disminución en las edades de 3, 4 y 5 años, posiblemente a causa de la situación generada por la pandemia del COVID-19.

Gráfico 1:

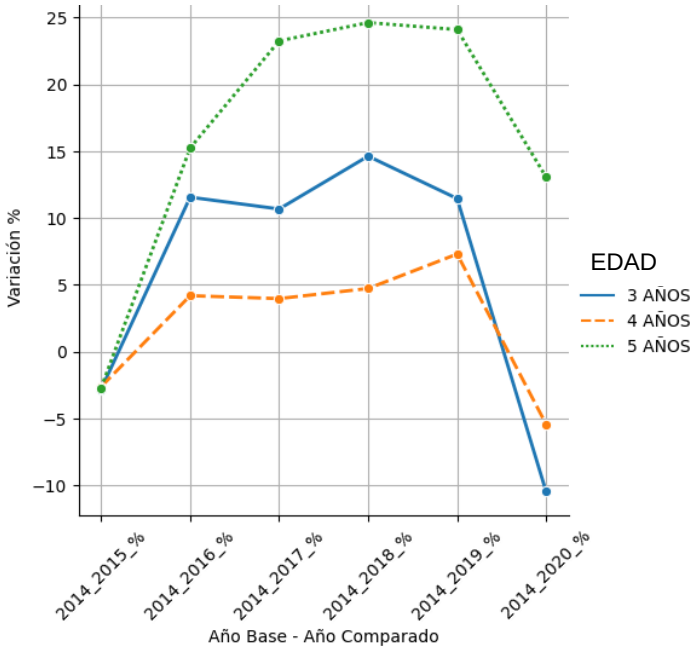
Evolución de la matrícula escolar del nivel inicial por edad.



En el Gráfico 2, se visualiza la variación porcentual en el nivel Inicial, utilizando las matrículas de 2014 como base para analizar el incremento porcentual anual. Se destaca un aumento más notable para la edad de «5 años», seguido de la edad de «3 años» y, finalmente, la edad de «4 años». Además, se observa una disminución con respecto al año 2020.

Gráfico 2:

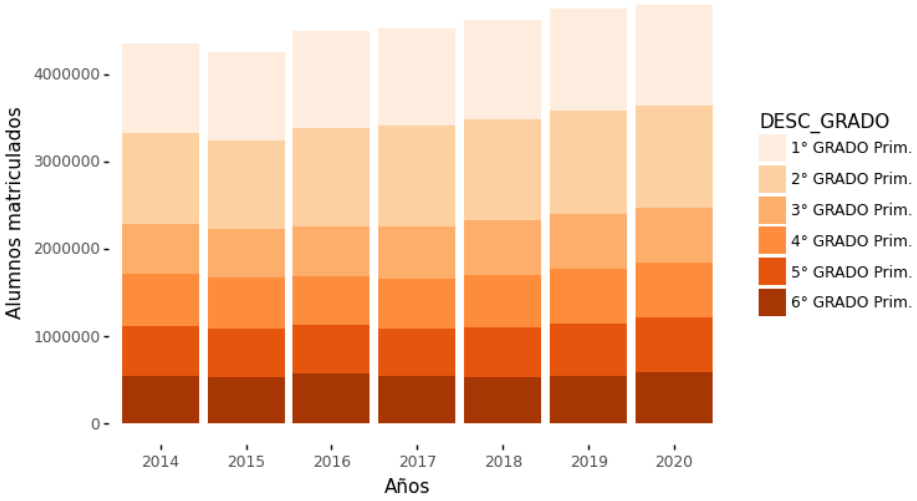
Variación % tomando como base al año 2014, nivel Inicial.



En el Gráfico 3, se visualiza la tendencia en el nivel Primaria. Se destaca un incremento desde el 1° Grado hasta el 4° Grado en todos los años, salvo por un leve descenso en todos los grados en el año 2015. Por otro lado, en los grados 5° y 6° de Primaria no se aprecia un aumento significativo en ningún año.

Gráfico 3:

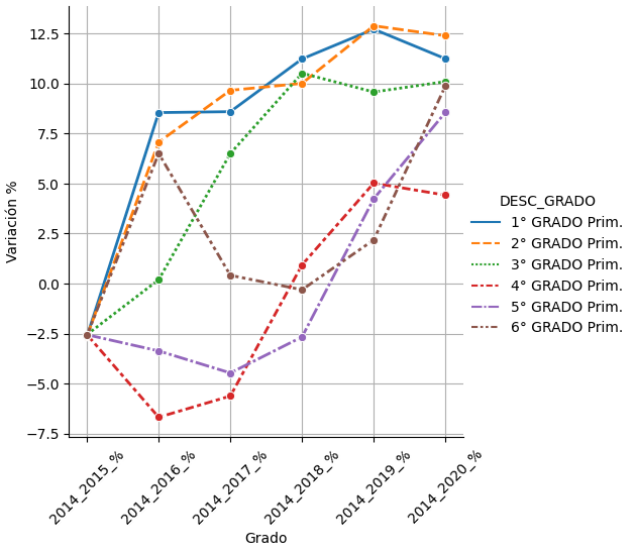
Evolución de la matrícula escolar del nivel primaria por grado.



En el Gráfico 4, se aprecia la variación porcentual en nivel Primaria, tomando las matrículas de 2014 como referencia para analizar el incremento porcentual anual. Se destaca un aumento más pronunciado en los grados de 1°, 2° y 3°. En contraste, los grados de 4°, 5° y 6° de Primaria experimentaron una disminución con tendencias posteriormente ascendentes.

Gráfico 4:

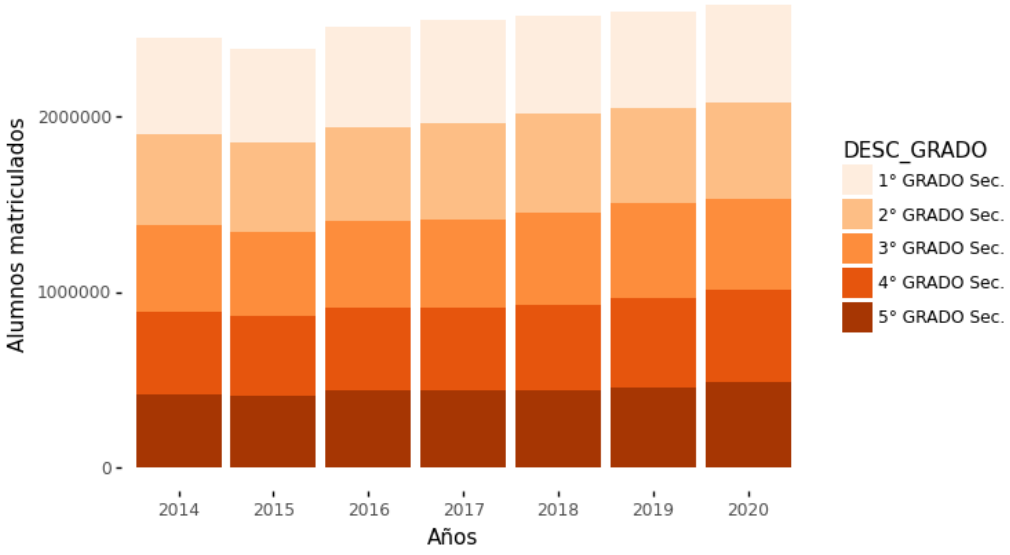
Variación % tomando como base al año 2014, nivel Primaria.



En el Gráfico 5, se evidencia la tendencia en el nivel de Secundaria. Se observa un aumento en la tendencia para todos los grados de educación secundaria (1°, 2°, 3°, 4°, 5°); sin embargo, como en casos anteriores, se aprecia una disminución en la tendencia durante el año 2015.

Gráfico 5:

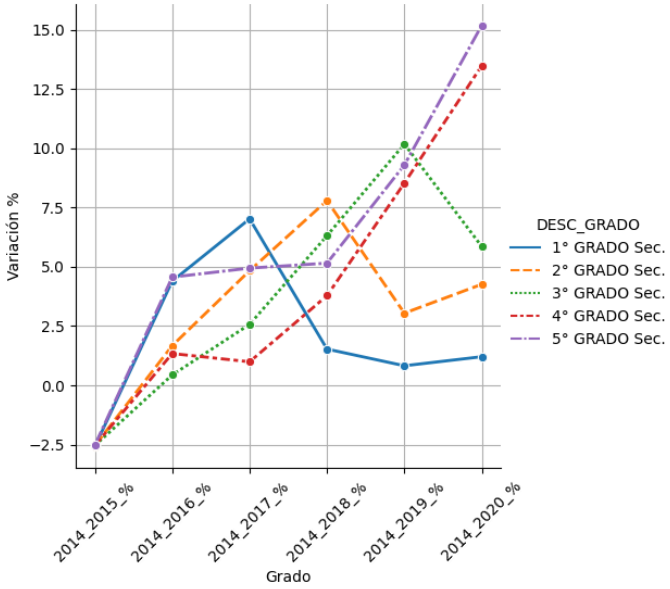
Evolución de la matrícula escolar del nivel secundaria por grado.



En el Gráfico 6, se aprecia la variación porcentual en el nivel Secundaria, tomando las matrículas de 2014 como referencia para analizar el incremento porcentual anual. Se destaca un aumento más pronunciado en los grados de 1°, 2° y 3°. En contraste, los grados de 4°y 5° de Secundaria experimentaron una disminución con tendencias posteriormente ascendentes.

Gráfico 6:

Variación % tomando como base al año 2014, nivel Secundaria.



2.4. VERIFICACIÓN: CALIDAD DE LOS DATOS

En esta etapa se revisan los datos bajo los siguientes puntos:

2.4.1. DATOS CON VALORES FALTANTES

La presencia reiterada de datos faltantes, expresados como valores nulos, constituye una problemática común en las fuentes de datos. Con el objetivo de enfrentar esta situación, se ha llevado a cabo una revisión exhaustiva de las fuentes, identificando este patrón en:

a. **SIAGIE**. En esta fuente de datos, se verificó que menos del 5% presenta valores nulos en columnas candidatas a ser variables. Como medida, se realizó la eliminación de los registros que contienen estos valores faltantes. Este procedimiento se repitió para los niveles Inicial, Primaria y Secundaria de los años 2018, 2019 y 2020.

b. **SINADEF**. Este conjunto de datos contiene algunos valores nulos, relacionados con el domicilio y su ubicación geográfica. Se procedió a crear una variable dicotómica adicional para poder identificarlos.

2.4.2. DATOS MÚLTIPLES

Los datos múltiples se refieren a registros que se presentan de manera múltiple en la misma fuente de datos o en distintas fuentes de datos. En la revisión de los datos se encuentra esta casuística en la siguiente fuente de datos:

a. **SINADEF**. Al realizarse la revisión de esta fuente de datos se encuentra algunos datos múltiples los cuales se proceden a quitarse.

2.4.3. ESTANDARIZACIÓN DE DATOS

El proceso de estandarización de los datos viene referido a las actividades de asignar valores estándares a distintas variables de distintas fuentes de datos, pero que comparten un significado común. Este proceso se lleva a cabo para poder realizar operaciones tales como segmentación, mezcla o depuración de los datos. Las fuentes de datos revisadas son:

a. **SIAGIE**. Para esta fuente de datos se realizan estandarizaciones de valores correspondientes a las descripciones de los grados, identificador de los grados, cálculo de edades promedio por grado y nivel académico.
b. **Censo Educativo**. Las estandarizaciones que se desarrollan para esta fuente de datos son de los valores sobre las descripciones de los grados, gestión de dependencia, modalidad de gestión, entre otros.
c. **SINADEF**. Al emplearse los datos del «Sistema Nacional de Defunciones», fue necesario estandarizar estos datos en cuanto a la denominación de la ubicación geográfica de cada defunción, los códigos de ubigeo[9] de domicilio y la descripción del tipo de defunción.
d. **Censo Nacional**. Los datos del Censo Nacional son estandarizados con los datos del Padrón Instituciones Educativas, generando variables derivadas en el proceso. Dichas estandarizaciones generan datos sobre distancias en kilómetros del censo educativo al punto central geográfico del ubigeo, variables sobre el área urbana o rural y variables sobre la pobreza.
e. **Población Estimada por Edades**. Estos datos al ser obtenidos de fuentes externas públicas (origen Ministerio de Salud) se tienen que vincularse con los datos del SIAGIE (año 2020), teniendo en cuenta el ubigeo y la edad según el grado al que pertenecen. Para lo cual se realizan estandarizaciones de dichas variables.

[9] Ubigeo: Código de ubicación geográfica provisto por INEI.

CAPÍTULO III: PREPARACIÓN DE LOS DATOS

3.1. SELECCIÓN DE DATOS

Esta etapa consiste selección de variables nativas a partir de las fuentes de datos identificadas. Para el presente informe, los datos son un conjunto de características etiquetadas que se pueden expresar matemáticamente como « $\{(x_i, y_i)\}_{i=1}^{N}$ », donde cada « x_i » de *N* es conocido como un vector de características. A su vez se puede decir que un vector de características abarca varias dimensiones, enumeradas desde j = 1, …, *D*, donde cada dimensión contiene un valor que se pretende representar. Este valor se denomina característica y es denotada por «$x^{(i)}$» (Burkov, 2022).

A partir de estas características, se llevó a cabo la selección de aquellas que podrían ser buenos predictores. Para lograr este objetivo, se llevan a cabo cinco procesos encargados de seleccionar características nativas, derivadas y agregadas. Estos procesos comprenden:

i. **Características de SIAGIE**. Los datos provenientes del SIAGIE proporcionan información relativa a características personales y temporales (anuales) relacionadas con la educación de los estudiantes. Como parte del análisis, se lleva a cabo la formación de un conjunto de caracteristicas agregadas por año, grado, sección y cantidad de estudiantes, considerando la edad media dentro de cada grupo respectivo. Estos conjuntos se crean de manera segmentada según el nivel educativo, abarcando los niveles Inicial, Primaria y Secundaria.
ii. **Características de Censo Educativo**. La selección de características se realiza en función de datos agregados, teniendo en consideración aspectos como el código modular, la cantidad de estudiantes, así como el año y grado al que estuvieron asociados.
iii. **Características de SINADEF**. Se generan características agregadas por año, código de provincia y cantidad de fallecidos.
iv. **Características de Censo Nacional**. Los datos del Censo Nacional son procesados conjuntamente con la información del Padrón de Instituciones Educativas de ESCALE. Este procesamiento tiene como resultado la creación de conjuntos de datos derivados que incorporan indicadores de ausentismo, analfabetismo y cantidades por nivel educativo. Además, se incluyen variables que indican la distancia en kilómetros desde la institución educativa hasta la ciudad o población de residencia del alumno.

v. **Características de Población Estimada por Edades**. Los datos obtenidos del portal web del Ministerio de Salud (MINSA) sobre las edades de la población estimada se combinan con los datos de SIAGIE, para obtener un nuevo conjunto de datos con variables adicionales con la población estimada.

3.2. PROCESAMIENTO DE LOS DATOS

Todas las fuentes de datos mencionados en el punto 3.1 fueron procesadas, bajo el siguiente esquema y procesos: 1) Limpieza de datos, 2) Construcción de nuevos datos, 3) Integración de datos y 4) Formato de datos.

Teniendo en cuenta este procesamiento se muestra la secuencialidad en el siguiente diagrama de procesos que se representa en la Ilustración 1.

i. **Entradas de datos**. Corresponde a la recopilación de distintas fuentes de datos sin realizar ningún tipo de tratamiento de datos. Las fuentes de datos corresponden al SIAGIE, Censo Educativo, Padrón de Instituciones Educativas, SINADEF, datos geográficos de códigos y distancias de ubigeo y poblaciones estimadas por el Censo Nacional del año 2017.

ii. **Procesamiento de datos**. En este paso se filtran, segmentan, refinan y equiparan los datos. Cada fuente de datos puede tener un tratamiento, pero adaptado a las variables que contiene. También se generan nuevas variables en base a agregaciones o combinación de distintas fuentes de datos. Cabe precisar que las tareas de imputación de datos se realizaron principalmente en la etapa final de procesamiento de datos (Salidas de datos).

iii. **Salidas de datos**. Esta sección abarca las salidas derivadas de las fases previas de procesamiento. Aquí, se extraen los conjuntos de datos resultantes que se utilizarán como insumo tanto para el entrenamiento del modelo así como la inferencia[10] para proyectar al 2021.

Es importante destacar que para recopilar los datos de matrícula, secciones y docentes, se utilizó principalmente el SIAGIE y el Censo Educativo. En la elección de la fuente de datos que proporcionaría información sobre estudiantes, secciones y docentes en cada grado de los servicios educativos, se optó por seleccionar la cantidad más alta entre ambas fuentes de información para dicho grado o sección. Si bien esta acción se aplicó de igual forma en los datos de entrenamiento e inferencia, es preciso señalar que se realizaron algunas transformaciones

[10] Los datos de inferencia constituyen el insumo para que el modelo estimado pueda proyectar la matrícula, secciones y docentes al año 2021.

específicas centradas únicamente en los datos de entrenamiento y otras en los datos de inferencia, las cuales se describen a continuación.

Preparación de los datos de entrenamiento:

Se organizaron en dos segmentos. En el primer segmento, se incorporaron los datos de matrícula del 2018 (dataframe18) como variables independientes, mientras que los datos de matrícula del 2019 (dataframe19) se establecieron como la variable dependiente. En el segundo segmento, los datos de matrícula del 2019 se emplearon como variables independientes, y los datos de matrícula del 2020 (dataframe20) se consideraron como la variable dependiente. Finalmente, se integraron ambos segmentos para generar los datos de entrenamiento, los cuales fueron guardados en el archivo master_data_matricula.csv. Es importante señalar que en el archivo master_data_matricula.csv también se han incluido las variables[11] asociadas a las fuentes de información descritas en la primera etapa del procesamiento de datos (Entrada de datos).

Preparación de los datos de inferencia:

Por otro lado, para formar los datos de inferencia (*master_predict.csv*), se utilizaron como variables independientes los datos de matrícula del 2020 (dataframe20) y las variables independientes asociadas a la fuente de información descritas en «Entrada de datos». Es relevante subrayar que en situaciones en las cuales no se contaba con información reciente sobre el número de estudiantes o secciones, se recurrió a registros históricos extendidos (2019 y 2018). En estos casos, se seleccionó el dato con el mayor valor para asegurar una estimación más precisa.

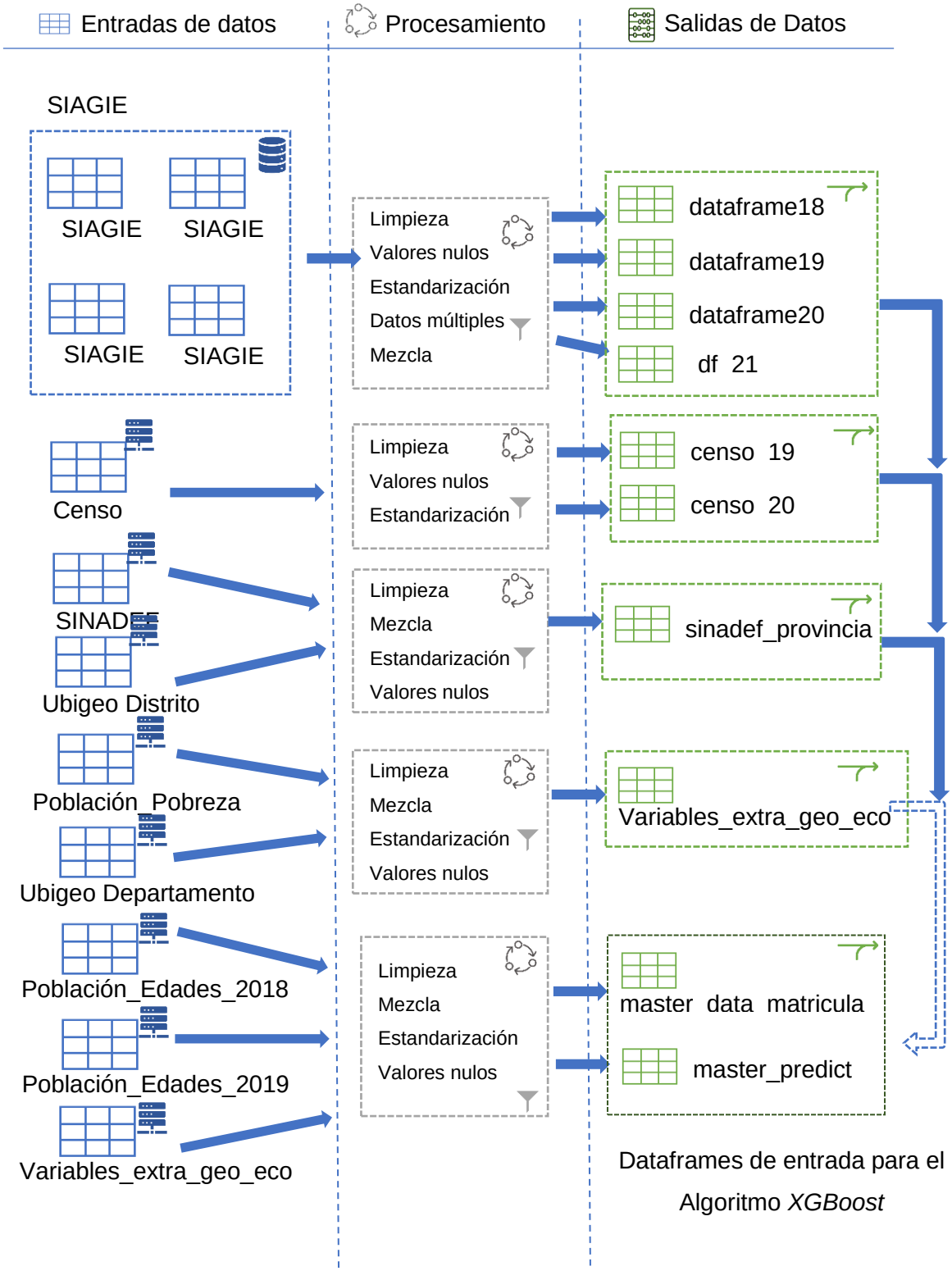
Asimismo, en el escenario de no contar con registros específicos para un grado en un nivel educativo de un servicio educativo, se asumió que el nivel inicial abarcaría las edades de 3, 4 y 5 años, mientras que para el nivel de primaria y secundaria se considerarán todos sus grados respectivos. Esta aproximación facilita una planificación educativa más acertada, garantizando la asignación adecuada de recursos y personal docente según las necesidades proyectadas.

Como resultado de la salida de datos, se crean los archivos *master_data_matricula.csv* y *master_predict.csv*, representando el resultado definitivo de todo el proceso de salida de datos.

[11] La descripción completa de todas las variables se encuentra detallada en la Tabla 4 del documento.

Ilustración 1:

Proceso de obtención de dataframes para el algoritmo XGBoost.



CAPÍTULO IV: MODELAMIENTO

En este capítulo, se proporcionan detalles sobre los criterios considerados y los resultados obtenidos durante la fase de modelado. Para ello, se emplearon los datos administrativos previamente descritos en capítulos anteriores, los cuales posibilitaron la estimación de un modelo robusto que proyecta la matrícula, las secciones y los docentes para el próximo año.

La estimación del modelo se basó principalmente en los datos de matrícula escolar de los años 2018, 2019 y 2020, así como en fuentes de información complementaria. Utilizando este estimador, se llevó a cabo un ejercicio de proyección para el año 2021 de la matrícula escolar, las secciones y los docentes. Estas proyecciones fueron comparadas con las realizadas por la UE para el mismo año. A continuación, se describen en detalle las acciones tomadas para llevar a cabo este proceso.

4.1. SELECCIÓN DE TÉCNICA DE MODELADO

En primer lugar, es esencial comprender qué implica el *Machine Learning*. Kaplan (2016) destaca que los programas de ML extraen patrones de los datos, constituyendo así el proceso central del aprendizaje de máquina. Además, señala que los datos pueden presentarse en diversas formas y variedades.

Russell (2018) indica que la decisión de cuándo utilizar ML se basa en consideraciones específicas. Cuando nos enfrentamos a problemas que requieren extensas listas de reglas para su resolución o cuando la complejidad de la tarea supera las capacidades de los enfoques tradicionales, se presenta una oportunidad para emplear ML. Esto se vuelve aún más relevante en entornos no estables que experimentan cambios constantes con la introducción de nuevos datos. En este contexto, la aplicación de técnicas avanzadas como el ML se convierte en una opción esencial para abordar eficientemente estos desafíos.

Para atender el objetivo de gestión que consiste en proyectar de manera más precisa la demanda educativa para un año futuro, se aborda el problema como un desafío de regresión. Este enfoque se justifica por la necesidad de determinar la cantidad futura de matrículas de alumnos, secciones y docentes, basándose en el comportamiento histórico. En otras palabras, se busca estimar una cantidad futura como variable objetivo.

Se ha optado por seleccionar el algoritmo «*eXtreme Gradient Boosting*» (XGBoost), reconocido habitualmente como un componente esencial en las soluciones triunfadoras en competiciones de ML (Géron, 2019). Para obtener una comprensión más completa

de este algoritmo, es esencial considerar los siguientes conceptos:

a. **Arboles de Decisión**: Un árbol de decisión es una estructura de datos jerárquica que implementa la estrategia divide y vencerás. Es un método no paramétrico eficiente, el cual puede ser utilizado para clasificación y regresión (Rokach & Maimon, 2008). Un árbol de decisión está compuesto de nodos de decisión interna y hojas terminales, donde cada nodo de decisión «m» implementa una prueba funcional «*fm(x)*» con resultados discretos que etiquetan las ramas.

b. ***XGBoost***[12]**:** Es un método de ML supervisado (Chen & Guestrin, 2016) para clasificación y regresión. Este método está basado en arboles de decisión. Es importante tener en cuenta las siguientes características de *XGBoost:*

   i. Es un ensamblado secuencial de árboles de decisiones. Se agregan arboles de decisión de manera secuencial con la finalidad de aprender del resultado de los árboles previos y corregir el error producido por los mismos, hasta lograr un mínimo error.

   ii. XGBoost emplea procesamiento en paralelo, poda de árboles, manejo de valores perdidos y regularización (optimización que penaliza la complejidad de los modelos) evitando de esta manera un sobreajuste del modelo.

   iii. Este algoritmo esta implementado como un paquete reusable de rutinas computacionales por la comunidad de miembros «xgboost developers», la cual le describen como: «XGBoost es una biblioteca de aumento de gradiente distribuida optimizada diseñada para ser altamente eficiente, flexible y portátil. Implementa algoritmos de Machine Learning bajo el marco Gradient Boosting» (xgboost developers, 2022).

Para afrontar el objetivo propuesto se hará uso de la librería ***XGBoost*** en su versión ***1.6.1***, específicamente el módulo ***XGBRegressor***.

[12] En el ANEXO 1 «PROCEDIMIENTO DE ESTIMACIÓN», se ofrece una descripción de los pasos que realiza dicho algoritmo en el proceso de estimación.

4.2. ESTRATEGIA DE MODELADO

En el «CAPÍTULO III, 3.1 SELECCIÓN DE DATOS», se definieron los grupos de variables pertinentes. A partir de estos grupos, se lograron identificar las variables que serán empleadas en el entrenamiento del modelo, las cuales se detallan en la Tabla 4.

Tabla 4:

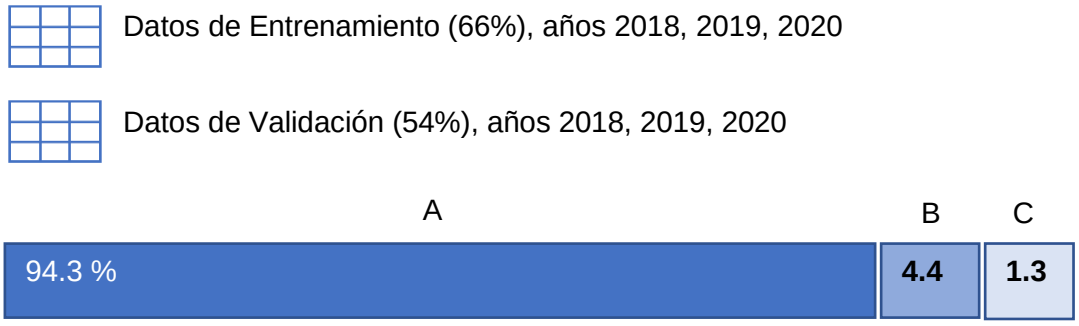
Variables para predicción de la matrícula escolar, secciones y docentes.

N°	Variable	Tipo	Descripción
1	max_est	Numérico	Variable objetivo con un horizonte temporal de un año: • Matrícula escolar. • Secciones / docentes.
2	GRADO	Numérico	Grado escolar de un nivel educativo de EBR. En el nivel inicial representa la edad.
3	EST	Numérico	Cantidad de estudiantes, teniendo en cuenta el grado y código modular al que pertenecen.
4	distance_km	Numérico	La distancia en kilómetros que se mide desde la ubicación del centro educativo (donde opera el código modular) hasta el punto central del ubigeo de la capital del departamento (suele ser la plaza principal).
5	POB_PROYECTADA	Numérico	Población proyectada a nivel distrital.
6	AUSENTISMO	Numérico	Indicador derivado de ausentismo a nivel distrital.
7	ANALFABETISMO	Numérico	Indicador derivado de analfabetismo a nivel distrital.
8	SECUNDARIA	Numérico	Población con secundaria completa a nivel distrital.
9	Fallecidos	Numérico	Número de fallecidos por provincia.
10	Edad_simple	Numérico	Número de alumnos dentro del segmento de edad para cursar un determinado grado por distrito.
11	SEC	Numérico	Número de secciones que se encuentran por grado (variable empleada solamente para el modelo de secciones y docentes).

En el proceso de entrenamiento del modelo, los datos se dividen en dos conjuntos: el grupo de entrenamiento y el grupo de validación. Ambos conjuntos poseen la misma cantidad de variables, aunque difieren en la cantidad de instancias. Dado que los datos abarcan dos años históricos (los segmentos correspondientes a los años 2018-2019 y 2019-2020), la separación se realiza de la siguiente manera:

Ilustración 2:

Segmentación de datos para el entrenamiento y validación.



(A) Servicios educativos con información histórica completa.

(B) Servicios educativos que no cuentan con información histórica en el periodo previo, pero sí para periodos anteriores.

(C) Servicios educativos nuevos que no cuentan con información histórica.

4.3. CONFIGURACIÓN DEL MODELO

4.3.1. CONFIGURACIÓN DEL ALGORITMO XGB-REGRESSOR

Un aspecto crucial a considerar en la generación de un modelo de *Machine Learning* son los hiperparámetros del algoritmo. Estos hiperparámetros, generalmente expresados como valores numéricos, aunque no exclusivamente, son propiedades que ejercen una influencia significativa en el resultado final del modelo (Burkov, 2022). Es importante destacar que los hiperparámetros no son aprendidos por el algoritmo; en cambio, deben ser especificados por el analista.

En el contexto del modelo de proyección de la demanda educativa, se ajustaron los hiperparámetros descritos en la tabla 5:

Tabla 5:

Definición de hiperparámetros empleados en XGBRegressor.

N	Hiperparámetros[13]	Definición
1	*nthread*	Cantidad de hilos paralelos empleados en la ejecución de *XGBoost*.
2	*objective*	Determina la tarea de aprendizaje y su respectivo objetivo.
3	*learning_rate*	Tasa de aprendizaje utilizada en el proceso de entrenamiento del modelo analítico con el fin de reducir el sobreajuste.
4	*max_depth*	Indica la profundidad máxima permitida para un árbol de decisión. Incrementar este valor incrementará la complejidad del modelo y su susceptibilidad al sobreajuste.
5	*min_child_weight*	Suma mínima del peso de instancia necesaria en los hijos.
6	*silent*	Controla la verbosidad de los mensajes de impresión.
7	*subsample*	Define la proporción de submuestra de instancias de entrenamiento.
8	*colsample_bytree*	Es la relación de submuestra de las columnas al construir cada árbol. El submuestreo ocurre una vez por cada árbol construido.
9	*n_estimators*	Regula la cantidad de árboles/estimadores en el modelo analítico. Valores elevados incrementan el riesgo de sobreajuste del modelo. Es crucial encontrar un equilibrio para obtener un rendimiento óptimo.

4.3.2. OPTIMIZACIÓN DE HIPERPARÁMETROS

La optimización de hiperparámetros se llevará a cabo utilizando el método RandomizedSearchCV[14]. Este método se encarga de buscar los valores óptimos de los hiperparámetros con el objetivo de mejorar el rendimiento del modelo de manera eficiente (scikit-learn, 2022). En la Tabla 6 se detallan las configuraciones empleadas de RandomizedSearchCV para la búsqueda de hiperparámetros.

[13] Descripciones obtenidas de https://xgboost.readthedocs.io/en/stable/parameter.html#

[14] Para mayor detalle de «RandomizedSearchCV» se recomienda revisar el portal https://scikit-learn.org.

Tabla 6:

Hiperparámetros para RandomizedSearchCV.

Hiperparámetros	Definición
parameters	Hiperparámetros de la tabla 5.
cv	Establece la cantidad de divisiones para la validación cruzada.
n_jobs	Cantidad de procesos en paralelo que se ejecutan.

4.4. GENERACIÓN DEL DISEÑO DE COMPROBACIÓN

Se definen como métricas de desempeño para los modelos el «Error Cuadrático Medio» (ECM o MSE, por sus siglas en inglés) y la «Raíz Cuadrada del Error Cuadrático Medio» (RMSE). Se prefiere el RMSE sobre el MSE debido a que el último puede amplificar los errores por ser una expresión cuadrática, haciendo al RMSE más intuitivo al estar en las mismas unidades que la variable de interés. Ambos valores se presentarán como las mediciones del error. El RMSE se utilizará para realizar comparaciones (variaciones porcentuales) entre los modelos generados y los modelos existentes de la UE. Es importante destacar que, según Armando Aguirre (1994), un RMSE menor indica una mayor adecuación del modelo. Se considera conveniente tener en cuenta los siguientes conceptos.

a. **ERROR CUADRÁTICO MEDIO (ECM o MSE en inglés).** Es un estimador que mide el promedio de los errores al cuadrado, es decir, muestra las diferencias entre el estimador y lo que se estimó. La diferencia es producida debido a la aleatoriedad de los datos o la falta de información para estimar de manera más precisa. También se puede enfatizar que el MSE es una forma de medir el error global que se da en la predicción (Heizer & Render, 2004). La fórmula para el cálculo es:

$$MSE \overset{\text{def}}{=} \frac{\sum(errores\ del\ pronóstico)}{n} = \frac{1}{n}\sum_{i=1}^{n}(\check{Y}_i - Y_i)^2$$

Donde:

$\check{Y}$: Es un vector de n predicciones.

Y: Es el vector de los valores verdaderos.

n: Número de observaciones.

b. **RAÍZ CUADRADA DEL ERROR CUADRÁTICO MEDIO (RECM O RMSE en inglés).** Esta expresión calcula la raíz cuadrada al MSE:

$$RMSE \overset{\text{def}}{=} \sqrt{\frac{\sum(errores\ del\ pronóstico)}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\check{Y}_i - Y_i)^2}$$

Donde:

$\check{Y}$: Es un vector de n predicciones.

Y: Es el vector de los valores verdaderos

n: Número de observaciones.

Con base a lo anterior, se calculó el MSE y RMSE para los modelos de proyección de la demanda educativa, con base a los datos reales del año 2021.

4.5. GENERACIÓN DE MODELOS

En línea con lo expuesto en el punto 4.1, se aplicó el algoritmo *XGBRegressor* para la estimación. En el ANEXO 1 «PROCEDIMIENTO DE ESTIMACIÓN», se ofrece una descripción detallada de los pasos que realiza dicho algoritmo en el proceso de estimación.

4.5.1. MODELOS DE MATRÍCULA

La obtención de los modelos de matrícula consta de las siguientes etapas:

a. **ETAPA I**. Para el entrenamiento de esta etapa se emplean los datos históricos, (representados por las variables del *Capítulo IV, 4.2. Estrategia de Modelado*), que representan el 94.3% (segmento A, véase *Ilustración 2*). De esta manera se obtiene el **Modelo I**, que proyecta la estimación de la matrícula escolar con base a todos los datos históricos completos. La tabla 7 muestra las variables empleadas:

Tabla 7:

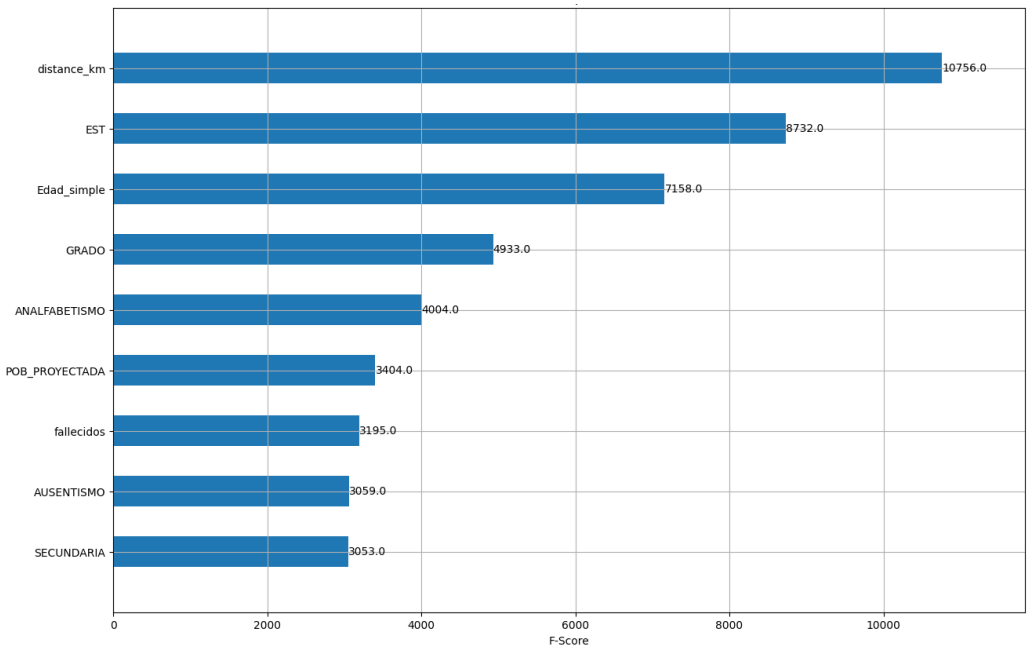
Variables para la estimación del Modelo I (proyección de matrícula).

Proceso	Variables	Descripción
Entrenamiento	i. GRADO ii. EST iii. distance_km iv. POB_PROYECTADA v. AUSENTISMO vi. ANALFABETISMO vii. SECUNDARIA viii. Edad_simple ix. fallecidos x. **max_est** (variable objetivo).	Ver Tabla 4.

Para determinar la importancia de las variables del modelo, se utiliza el F-Score[15]. Un valor más alto del F-Score para una variable está directamente relacionado con su capacidad predictiva en el modelo.

Gráfico 7:

Importancia de las variables en el Modelo I.



El Gráfico 7 presenta la importancia de las variables medido por el F-Score. Se puede apreciar que la variable *distance_km* es la que más información aporta al modelo, seguido de la variable *EST* que representa la cantidad de alumnos por código modular y grado (en años anteriores). En tercer lugar, se tiene el predictor *Edad_simple* que representa la cantidad de población estimada por edad por código geográfico para un determinado grado.

b. **ETAPA II**. Para esta etapa se ejecuta la predicción con el modelo que se obtuvo en la ETAPA I, pero solo considerando el 4.4% de los datos (segmento B, ver *Ilustración 2*). La variable *EST,* al ser un predictor requerido por el modelo (pero sin datos para el segmento B), requiere de una imputación empleando la media de los años 2018, 2019, 2020. La tabla 8 muestra las variables empleadas en esta etapa:

[15] El F-score es una métrica que se basa en la frecuencia con la que una variable aparece en los árboles de decisión del modelo. Cuanto más se emplea una variable para tomar decisiones en los nodos de los árboles de decisiones del modelo, mayor importancia tendrá esa variable.

Tabla 8:

Variables para predicción con Modelo I (proyección de matrícula).

Proceso	Variables	Descripción
Predicción	i. GRADO ii. EST iii. distance_km iv. POB_PROYECTADA v. AUSENTISMO vi. ANALFABETISMO vii. SECUNDARIA viii. Edad_simple ix. fallecidos x. **max_est** (variable objetivo).	Ver Tabla 4.

c. **ETAPA III**. Esta última etapa se desarrolla para los códigos modulares que son completamente nuevos para el año 2020, es decir códigos modulares que no tienen ningún tipo de dato histórico. El segmento de datos es representado por el 1.3% (segmento C, ver *Ilustración 2*). No se considera la columna *EST*. Como resultado de este proceso se obtiene el **Modelo II**. Las variables empleadas para el modelo II son:

Tabla 9:

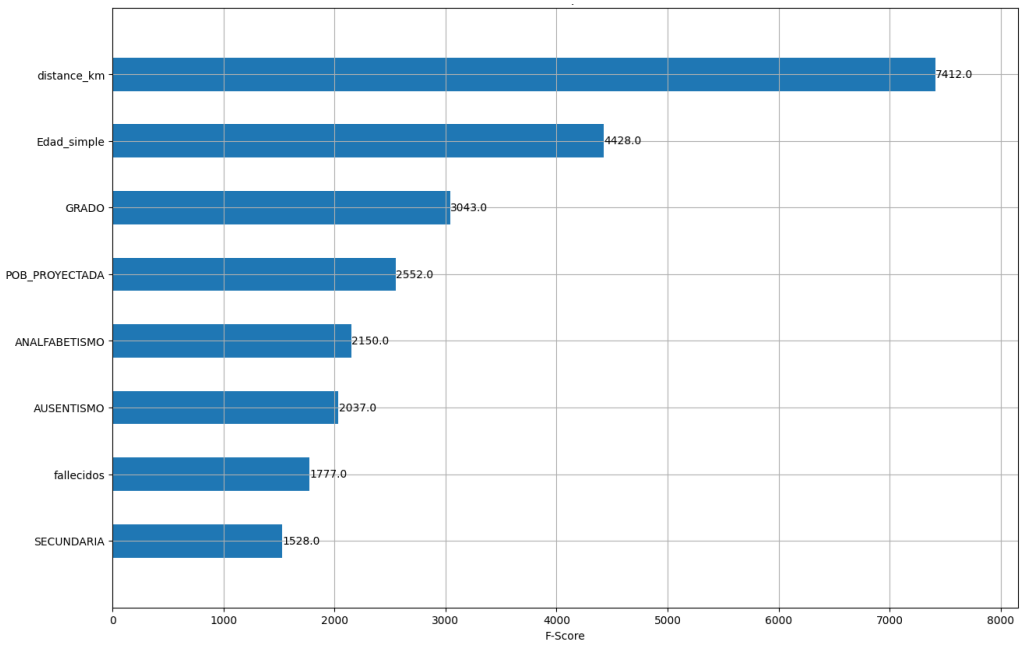
Variables para estimación del Modelo II (proyección de matrícula).

Proceso	Variables	Descripción
Entrenamiento	i. GRADO ii. distance_km iii. POB_PROYECTADA iv. AUSENTISMO v. ANALFABETISMO vi. SECUNDARIA vii. Edad_simple viii. fallecidos ix. **max_est** (variable objetivo).	Ver Tabla 4.

En el Gráfico 8 se evidencia que también la variable *distance_km*, sigue siendo la más importante, seguido de la *Edad_Simple*, *grado* y *POB_PROYECTADA*.

Gráfico 8:

Importancia de Variables para el Modelo II.



4.5.2. MODELOS DE SECCIONES Y DOCENTES

ETAPA I. Esta etapa también emplea los datos históricos completos (segmento A con 94.3%, ver *Ilustración 2*). A diferencia del modelo de proyección de matrícula, para este caso se emplea adicionalmente la variable *SEC* (Número de secciones por grado). Como resultado se obtiene el **Modelo III**, que proyecta la estimación de secciones y docentes y emplea las siguientes variables:

Tabla 10:

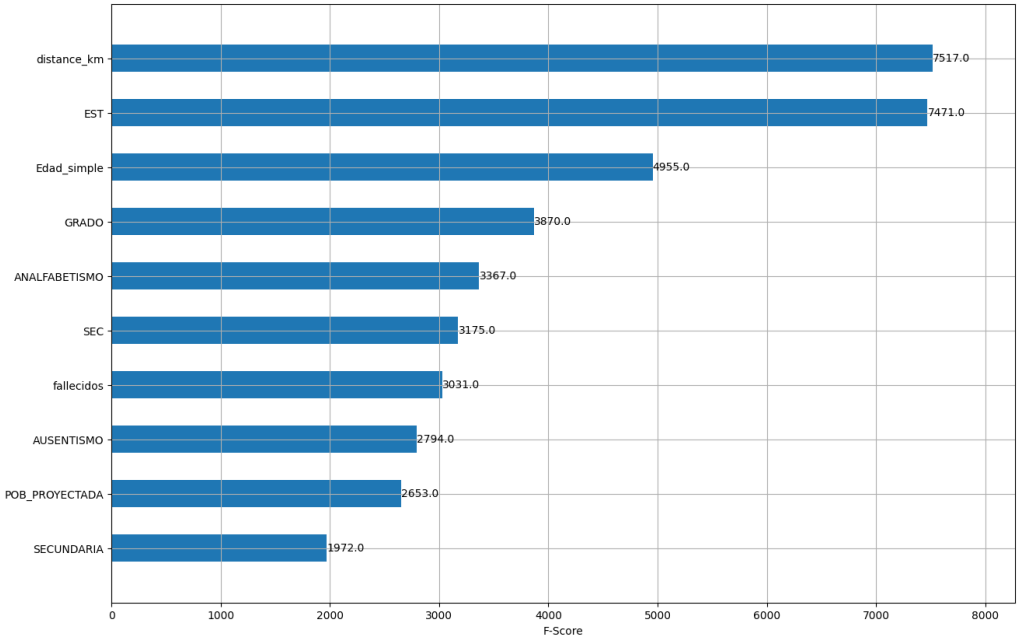
Variables para estimación del Modelo III (secciones y docentes).

Proceso	Variables	Descripción
Entrenamiento	i. GRADO ii. EST iii. SEC iv. distance_km v. POB_PROYECTADA vi. AUSENTISMO vii. ANALFABETISMO viii. SECUNDARIA ix. Edad_simple x. fallecidos xi. **max_est** (variable objetivo).	Ver Tabla 4.

El *Gráfico 9* presenta la importancia de las variables medido por el F-Score. Se aprecia que la variable *distance_km* se mantiene como la más importante, seguido *EST* y Edad_simple.

Gráfico 9:

Importancia de Variables para el Modelo III.



a. **ETAPA II**. Esta etapa es similar al modelo I de proyección de matrícula. Aquí se ejecuta la predicción con el modelo que se obtuvo en la ETAPA I, considerando solamente el 4.4% de los datos (segmento B, ver *Ilustración 2*).

Tabla 11:

Variables para predicción con el Modelo III (secciones y docentes).

Proceso	Variables	Descripción
Predicción	i. GRADO ii. EST iii. SEC iv. distance_km v. POB_PROYECTADA vi. AUSENTISMO vii. ANALFABETISMO viii. SECUNDARIA ix. Edad_simple x. fallecidos xi. **max_est** (variable objetivo).	Ver Tabla 4.

b. **ETAPA III**. Realizado para los códigos modulares que son completamente nuevos para el año 2020 (códigos modulares que no poseen ningún tipo de dato histórico). El segmento de datos es representado por el 1.3% (segmento C, véase *Ilustración 2*). Tampoco se considera la columna *EST*. Resultado de este proceso se obtiene el **Modelo IV**. Las variables empleadas son:

Tabla 12:

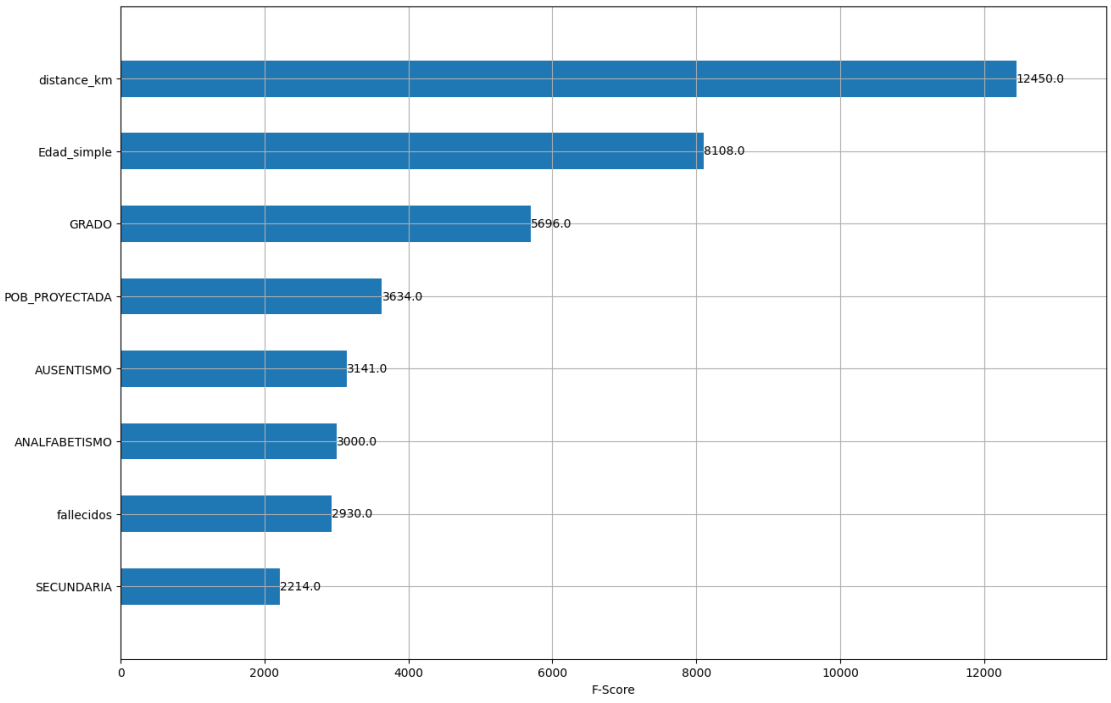
Variables para estimación del Modelo IV (secciones y docentes).

Proceso	Variables	Descripción
Entrenamiento	i. GRADO ii. distance_km iii. POB_PROYECTADA iv. AUSENTISMO v. ANALFABETISMO vi. SECUNDARIA vii. Edad_simple viii. fallecidos ix. **max_est** (variable objetivo).	Ver Tabla 4.

El Gráfico 10 muestra las variables que más aportan en el modelo.

Gráfico 10:

Importancia de Variables para el Modelo IV.



En el Gráfico 10 se nota que también para este caso la variable *distance_km* sigue siendo la que más información aporta (mayor valor F-Score), la *Edad_Simple* es el segundo predictor más importante, seguido del *grado* y *POB_PROYECTADA*. Pero no se considera la variable *SEC,* ni la variable EST ya que esta etapa corresponde a grados y secciones que no cuenta con ningún dato histórico.

4.6. EVALUACIÓN DE LOS MODELOS

Se procede al cálculo del MSE y RMSE que se realizará para los modelos I, II (proyección de matrícula) y modelos III, IV (proyección de secciones y docentes).

El cálculo se realizará en comparación con los matriculados reales de los niveles de Inicial, Primaria y Secundaria de EBR para el año 2021 (registrados en el SIAGIE), como también con las secciones y docentes registrados para dicho año. Es necesario mencionar que los resultados de los modelos I y II (proyección de matrículas) son complementarios, lo mismo ocurre con los resultados de los modelos III y IV (proyección de secciones y docentes).

El cálculo de la MSE y RMSE para la proyección de matrículas es como sigue:

Tabla 13:

MSE y RMSE para modelo I y II: matrícula - XGBRegressor.

Métrica	Valor
MSE	30.04
RMSE	5.48

También se calcula el MSE y RMSE para los modelos III, IV (modelo de proyección de secciones y docentes), la comparación se dará con la cantidad de secciones encontradas para el año 2021 del SIAGIE.

Tabla 14:

MSE y RMSE para modelo III y IV: secciones y docentes.

Métrica	Valor
MSE	0.08
RMSE	0.28

Se calcula el MSE y el RMSE para la proyección de matrícula realizada por la metodología de la UE, proyección de matrícula para el año 2021 modalidad de Educación Básica Regular y niveles Inicial, Primaria y Secundaria. Teniendo los siguientes resultados:

Tabla 15:

MSE y RMSE para matrícula (UE).

Métrica	Valor
MSE	64.06
RMSE	8.00

De similar manera se calcula el MSE y el RMSE para la proyección de secciones y docentes bajo la metodología de la UE. Obteniendo lo siguiente:

Tabla 16:

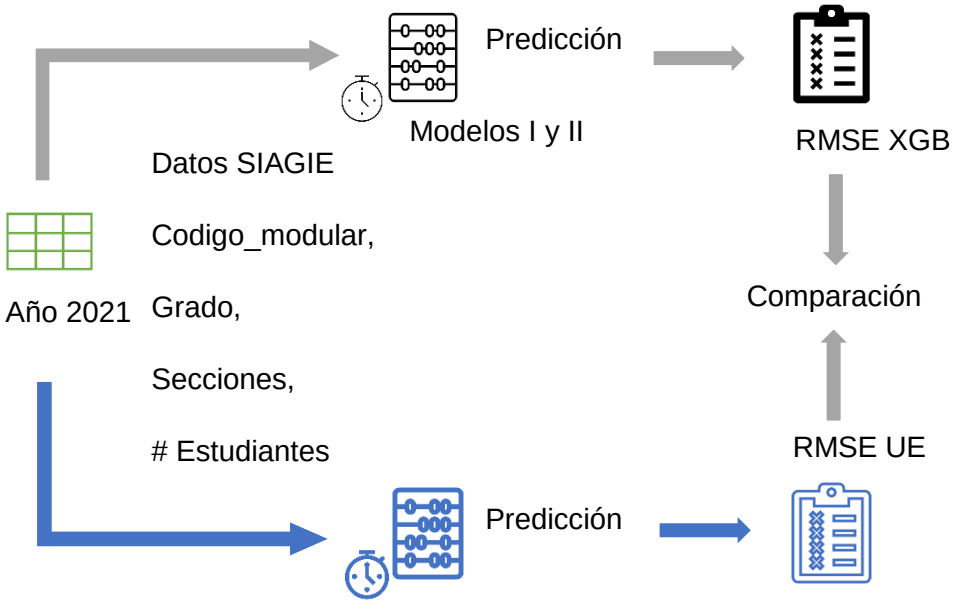
MSE y RMSE para secciones y docentes (UE).

Métrica	Valor
MSE	0.14
RMSE	0.37

Para entender mejor el proceso de comparación de los modelos I y II (*XGBRegressor*) versus la proyección realizada por la UE del Minedu, se muestra la siguiente ilustración donde se toma los datos del SIAGIE, específicamente los matriculados del año 2021 (Código modular, grado, cantidad de estudiantes) como base para el cálculo de los RMSE (Modelo I, II y Proyección de la UE), para luego comparar los resultados.

Ilustración 3:

Comparación del RMSE para los Modelos I y II y el modelo de la UE.



Según la comparación de resultados, el RMSE del modelo combinado I y II obtiene el menor valor:

$$5.48_{\substack{RMSE\ Modelo \\ I\ y\ II}} < 8.00\ _{\substack{RMSE\ Proyección \\ matrícula\ UE}}$$

Se podría evidenciar que existe cierta mejora el cual se examinará en el capitulo.

Para los modelos III, IV (Proyección de secciones y docentes) se sigue el mismo proceso de la *Ilustración 3,* pero tomando como referencia a las secciones por código modular y grado, la comparación de los modelos generados (III, IV) y modelo de UE también se realiza en base a datos reales de secciones registradas en el SIAGIE para el año 2021. Se encuentra el resultado siguiente:

$$0.29\ _{\substack{RMSE \\ Modelo \\ III\ y\ IV}} < 0.37\ _{\substack{RMSE\ Proyección \\ secciones\ y\ docentes \\ UE}}$$

CAPÍTULO V: EVALUACIÓN

5.1. COMPARACIÓN CON METODOLOGÍAS

La comparación de los modelos se realiza para los niveles Inicial, Primaria, Secundaria de EBR agrupado por grado para los modelos I y II (proyección de matrícula) y modelos III y IV (proyección de secciones y docentes).

En primer lugar, se procedió a calcular el RMSE correspondientes a la proyección de matrícula para el año 2021, para la nueva metodología de proyección de matrícula (modelos I y II) y la actual metodología de la UE. Posteriormente, se llevaron a cabo los cálculos de las diferencias porcentuales entre los RMSE de ambas metodologías.

De manera similar, se efectuaron los cálculos del RMSE para la proyección de secciones y docentes en el año 2021, para la nueva metodología de proyección de secciones y docentes (modelos III y IV) y la metodología de la UE. Seguidamente, se realizaron los cálculos correspondientes a las diferencias porcentuales de ambas metodologías.

Asimismo, los modelos III y IV de proyección de secciones (modelos obtenidos en Etapas III y IV) es equivalente a la aproximación de cantidad de docentes para cada código modular, grado y sección. Este enfoque sigue el criterio de la UE, detallado en la Etapa 1, referencia 9. (Minedu, 2022b).

5.1.1. EVALUACIÓN: MATRÍCULA NIVEL INICIAL

La Tabla 17 resalta una diferencia porcentual positiva en el RMSE para el modelo I y II en el nivel Inicial, con un 10.93% para la edad de «3 años» y un 31.25% para la edad de «5 años». Estos hallazgos evidencian una mayor precisión en el rendimiento del modelo I y II para estas edades. Por otro lado, en el caso de la edad de «4 años», se observa una diferencia porcentual negativa en el RMSE de -17.46%, indicando que el modelo de la UE exhibe un desempeño superior al modelo I y II.

Tabla 17:

Diferencia porcentual del RMSE: matrícula - nivel Inicial.

Edad	Modelo UE	Modelo I y II (XGBRe)[16].	Diferencia %
3 años	6.40	5.70	10.94
4 años	6.30	7.40	**-17.46**
5 años	6.40	4.40	31.25

[16] Acrónimo del algoritmo XGBRegressor basado en la técnica XGBoost.

5.1.2. EVALUACIÓN: MATRÍCULA NIVEL PRIMARIA

La Tabla 18 destaca una diferencia porcentual positiva en el RMSE para el modelo I y II en el nivel de Primaria de EBR, fluctuando entre un rango de 18.18% y 47.06%. Estos resultados subrayan una mayor precisión en el desempeño del modelo I y II en todos los grados de dicho nivel en comparación con el modelo actual de la UE.

Tabla 18:

Diferencia porcentual del RMSE: matrícula - nivel Primaria.

Grado	Modelo UE	Modelo I-II (XGBRe)	Diferencia %
1°	7.7	6.3	18.18
2°	7.7	4.3	44.15
3°	6.4	4.0	37.50
4°	6.8	3.6	47.06
5°	7.1	3.9	45.07
6°	5.4	3.6	33.33

5.1.3. EVALUACIÓN: MATRÍCULA NIVEL SECUNDARIA

La Tabla 19 destaca una diferencia porcentual positiva en el RMSE para el modelo I y II en el nivel de Secundaria de EBR, fluctuando entre un rango de 22.72% y 58.09%. Estos resultados subrayan una mayor precisión en el desempeño del modelo I y II en todos los grados de dicho nivel en comparación con el modelo actual de la UE.

Tabla 19:

Diferencia porcentual del RMSE: matrícula - nivel Secundaria.

Grado	Modelo UE	Modelo III-IV (XGBRe)	Diferencia %
1°	15.4	11.9	22.72
2°	12.5	6.3	49.60
3°	12.5	5.8	53.60
4°	13.6	5.7	58.09
5°	10.2	5.4	47.06

5.1.4. EVALUACIÓN: SECCIONES Y DOCENTES NIVEL INICIAL

La Tabla 20 resalta un aumento porcentual positivo en el RMSE para el modelo III y IV, los cuales realizan proyecciones de secciones y docentes en el nivel Inicial de EBR. Esta mejora se sitúa en un intervalo que varía entre el 25% y el 50%. Estos resultados indican que el modelo III y IV exhibe una precisión superior para todas las edades en comparación con el modelo actual de la UE. Sin embargo, es necesario mencionar que las diferencias entre los RMSE de los modelos están en el orden decimal, es decir no se aprecia gran diferencia.

Tabla 20:

Diferencia porcentual del RMSE: seciones y docentes - nivel inicial.

Edad	Modelo UE	Modelo III-IV (XGBRe)	Diferencia %
3 años	0.3	0.2	33.33
4 años	0.4	0.3	25.00
5 años	0.4	0.2	50.00

5.1.5. EVALUACIÓN: SECCIONES Y DOCENTES NIVEL PRIMARIA

La Tabla 21 muestra que, para el nivel primaria, la diferencia porcentual de los RMSE del modelo III y IV en contraste con el modelo de secciones y docentes de la UE no presenta ninguna diferencia a excepción del 6° grado, donde el modelo del UE es superior (menor RMSE). Los RMSE de ambos modelos están en el orden decimal. Sin embargo, es necesario mencionar que las diferencias entre los RMSE de los modelos están en el orden decimal, es decir no se aprecia gran diferencia.

Tabla 21:

Diferencia porcentual del RMSE: secciones y docentes - nivel primaria.

Grado	Modelo UE	Modelo III-IV (XGBRe)	Diferencia %
1°	0.3	0.3	0.0
2°	0.3	0.3	0.0
3°	0.3	0.3	0.0
4°	0.3	0.3	0.0
5°	0.3	0.3	0.0
6°	0.2	0.3	-50.0

5.1.6. EVALUACIÓN: SECCIONES Y DOCENTES NIVEL SECUNDARIA

La Tabla 22 muestra que, en el nivel secundaria, la diferencia porcentual de los RMSE de los modelos III y IV y el modelo de secciones y docentes de la UE presenta una diferencia constante del 33%, en favor de los modelos III y IV. Sin embargo, esta diferencia no es significativa.

Tabla 22:

Diferencia porcentual del RMSE: secciones y docentes - nivel secundaria.

Grado	Modelo UE	Modelo I-II (XGBRe)	Diferencia %
1°	0.6	0.4	33.33
2°	0.6	0.4	33.33
3°	0.6	0.4	33.33
4°	0.6	0.4	33.33
5°	0.6	0.4	33.33

CONCLUSIONES

Mediante el presente informe se describe la metodología empleada para desarrollar los modelos de Machine Learning (ML) destinado a proyectar, con un horizonte temporal de un año, la matrícula escolar, las secciones y los docentes correspondientes a cada grado en los servicios educativos de la Educación Básica Regular (EBR) en el Perú. Este hito refleja el logro exitoso del objetivo específico «A» que se detalla en este informe.

Los modelos de ML desarrollados se fundamentan en una regresión basada en el algoritmo *XGBoost*. Este método se destaca por su habilidad esencial para identificar y modelar relaciones no lineales eficazmente. Dicha capacidad es fundamental, ya que permite al modelo comprender cómo las variables de entrada interactúan entre sí y cómo estas influencias se relacionan directamente con la variable de salida. Al hacerlo, el modelo puede predecir resultados con mayor precisión, considerando la complejidad y dinamismo real de los datos. La elección de este diseño algorítmico se alinea directamente con el objetivo específico «B».

La evaluación de los resultados obtenidos del modelo de ML para la proyección de la matrícula escolar revela una reducción significativa del Error Cuadrático Medio de la Raíz (RMSE) en comparación con el modelo actual de la Unidad Educativa (UE). Estas mejoras son notables en los niveles de Primaria y Secundaria de la EBR en el Perú. En términos generales, se observa que el promedio de las diferencias porcentuales del RMSE en todos los grados del nivel Primaria es del 37.54%, mientras que en el nivel Secundaria es del 46.21%. Además, en el nivel Inicial se aprecia una mejora en la diferencia porcentual promedio del 8.24%. Cabe destacar que, para el nivel Inicial de 4 años, la diferencia porcentual fue de -17.46%, indicando una superioridad del modelo de la UE para este caso particular.

Adicionalmente, los modelos de ML utilizados tanto para la proyección de secciones como para la proyección de docentes muestran cierta mejora en comparación con el modelo actual de la UE. Sin embargo, es importante destacar que esta mejora no alcanza a ser significativa.En consecuencia, se puede afirmar que los resultados de la evaluación logran evidenciar que se ha cumplido con el objetivo específico «C», el cual buscaba reducir el RMSE por debajo del nivel actual en el modelo utilizado por la UE.

A partir de la atención de los objetivos específicos, podemos concluir que la metodología propuesta en este informe técnico logró cumplir el objetivo de la gestión, el cual buscó mejorar la precisión en la proyección de la matrícula escolar, las secciones y los docentes de los servicios educativos de EBR en el Perú.

LÍNEAS DE MEJORA

Como futura mejora a la metodología se tiene contemplado realizar una exploración proactiva de nuevas variables espaciales como un enfoque estratégico para enriquecer sustancialmente los modelos, ya que se pudo evidenciar que variables como «distance_km» fueron muy importantes en la estimación del modelo. Esta estrategia no solo representa una oportunidad para optimizar la precisión del modelo, sino que también posibilita la reducción significativa del RMSE en las proyecciones de la demanda educativa.

Otra oportunidad de mejora está relacionada con realizar una comparación exhaustiva con otros algoritmos, constituyendo así la segunda iteración de este estudio. Se plantea explorar la aplicación del aprendizaje combinado, una estrategia que implica la formación de varios modelos de baja precisión con el fin de construir un metamodelo de alta precisión. El propósito subyacente radica en evaluar minuciosamente si el rendimiento del algoritmo actual (*XGBoost-XGBRegressor*) supera a sus contrapartes en términos de métricas específicas. Este enfoque permitirá determinar con mayor claridad la efectividad de la elección algorítmica.

En ciertos contextos, resulta esencial disponer de una proyección a dos años para la matrícula escolar, secciones y docentes. Esta nueva perspectiva implicará realizar nuevos ajustes al modelo actual y evaluar las nuevas métricas de rendimiento.

Por otro lado, se tiene contemplado llevar a cabo un análisis más detallado sobre la sobreestimación y subestimación en cada uno de los grados de EBR. Este enfoque permitirá identificar las razones subyacentes detrás de estos errores, con el objetivo de mejorar la precisión del modelo.

Por último, es importante realizar un monitoreo continuo de las fuentes de información, las métricas de rendimiento y los cambios de contexto previamente analizados en este informe. Esto facilitará la detección de oportunidades para mejorar la calidad de la información utilizada y mejorar las métricas de robustez obtenidas. Además, permitirá alertar ante cualquier necesidad de ajuste que pueda requerir los modelos ML para un año escolar específico. De esta manera, se busca lograr una respuesta más eficaz y adaptada a las demandas cambiantes de la gestión educativa.

BIBLIOGRAFÍA

Ahmad, S., Don, H., Tushar, O., & Terry, B. (2018). Predicting Student Enrollment Based on Student and College Characteristics. *International Educational Data Mining Society*.

Armando Aguirre, J. (1994). *Introducción al Tratamiento de Series Temporales: Aplicación a las Ciencias de la Salud.* Madrid: Díaz de Santos S.A.

Armstrong, D. F., & Nunley, C. W. (1981). Enrollment Projection Within a Decision-Making Framework. *The Journal of Higher Education*.

Burkov, A. (2022). *The Hundred Page Machine Learning.* Themlbook.

Candela Rojas, E. C., & Centeno Guzmán, C. D. (2022). Alerta Escuela: Machine Learning para el cálculo del riesgo de interrupción de estudios en el Perú.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. 785-794.

Deng, H., Zhou, Y., Wang, L., & Zhang, C. (2021). *Ensemble learning for the early prediction of neonatal jaundice with genetic features.* BMC medical informatics and decision making.

Egbo, M. N., & Bartholomew, D. C. (2018). Forecasting Students' Enrollment Using Neural Networks and Ordinary Least Squares Regression Models. *Journal of Advanced Statistics*.

Fraysier, K., Amy, R., & James, A. (2020). Predicting Postsecondary Enrollment With Secondary Student Engagement Data. *Psychoeducational Assessment*, 1-18.

Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine.*

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow.* O'REILLY.

Heizer, J., & Render, B. (2004). *Principios De Administración de Operaciones.* México: Pearson Education.

Hussar, W. J., & Bailey, T. M. (2011). Projections of Education Statistics to 2020. *National Center for Education Statistics*.

James, F. E., & Weese, J. L. (2022). Neural Network-Based Forecasting of Student Enrollment With Exponential Smoothing Baseline and Performance Analysis. *In*

2022 ASEE Annual Conference & Exposition.

Kaplan, J. (2016). *Artificial Intelligence.* Oxford: Oxford University Press.

Kovačić, Z. (2010). Early Prediction of Student Success: Mining Students Enrolment Data. *Proceedings of Informing Science & IT Education Conference (InSITE)* (pág. 19). Wellington: Open Polytechnic.

Minedu. (9 de Febrero de 2022a). Norma sobre el proceso de matrícula en la Educación Básica. Lima, Lima, Perú.

Minedu. (2022b, Agosto 1). INFORME N° 00070-2022-MINEDU/SPE-OSEE-UE. Lima, Perú.

Pérez, C. (2021). *DATA MINING. The CRISP-DM Methodology. The CLEM language and IBM SPSS MODELER.* Scientific Books .

Rana, A., Iqbal, N., & Saima, T. (2015). The influence of Parents Educational level on Secondary School Students Academic achievemts in District Rajanpur. *Education and Practice*.

Reichardt, R., Klute, M., Joshua, S., & Stephen, M. (2020). *An Approach to Using Student and Teacher Data to Understand and Predict Teacher Shortages.* EE.UU: U.S. Department of Education.

Research Division, Nacional Education Association. (1953). Teacher Forecast For The Public Schools. 53-58.

Research Division, National Education Association. (1953). The Outlook for School Enrollments: Research Division, National Education Association. *Journal of Teacher Education*, 46-52.

Rokach, L., & Maimon, O. (2008). *Data Mining with Desicion Trees.* Singapore: World Scientific Publishing Co. Pte. Ltd.

Russell, R. (2018). *Machine Learning.*

scikit-learn. (11 de 10 de 2022). *sklearn.model_selection.RandomizedSearchCV* ¶. Obtenido de https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

Smelser, N. J., & Baltes, P. B. (2001). International Encyclopedia of the Social & Behavioral Sciences. *Elsevier*.

Wright, J. D. (2015). International Encyclopedia of the Social & Behavioral Sciences (Second Edition).

xgboost developers. (7 de Octubre de 2022). *dmlc XGBoost*. Obtenido de https://xgboost.readthedocs.io/en/stable/contrib/index.html

Yang, L., Tian, L., Yu, L., & Chen, Y. (2020). Research and Analysis on the Prediction of College Erollment based on Random Forest. *2020 International Conference on Education, Sport an Psychological Studies* (pág. 6). Dongguan: ESPS.

Zhuang, Y., & Gan, Z. (2017). A machine learning approach to enrollment prediction in Chicago Public School. *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*.

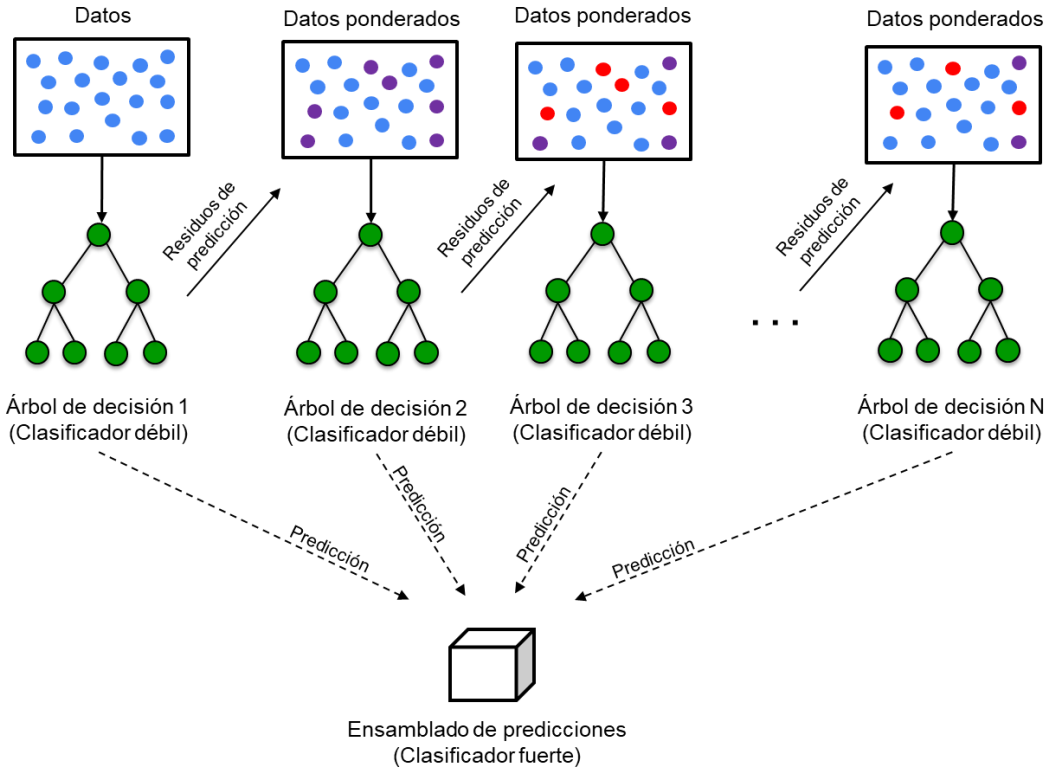
ANEXO 1: PROCEDIMIENTO DE ESTIMACIÓN

Para entender la arquitectura del *XGBoost*, nos enfocaremos en la estructura del algoritmo de *Gradient Boosting Decision Tree*[17] (GBDT), ya que XGBoost se construye como una implementación de este. Este enfoque nos brindará la oportunidad de simplificar teóricamente el proceso mediante el cual XGBoost realiza sus predicciones.

La Ilustración 4 tiene como objetivo representar cómo se lleva a cabo la estimación de las predicciones mediante el algoritmo de GBDT.

Ilustración 4:

Arquitectura del algoritmo de Gradient Boosting Decision Tree.



Nota. Adaptado de *The architecture of Gradient Boosting Decision Tree* [Figura], por Deng H. et al., (2021).

A continuación, se describen los procesos descritos en la Ilustración 4, los cuales son clave para la estimación de predicciones mediante el algoritmo de GBDT:

1. **Primer árbol:** Se entrena el «Árbol de decisión 1» para predecir los valores reales de la variable objetivo que se desea proyectar. Luego de esto, se procede a calcular los residuos, que representan las diferencias entre las predicciones actuales y los

[17] Para mayor detalle del modelo GBDT, se recomienda revisar «*Greedy function approximation: A gradient boosting machine*» (Friedman, 2001).

valores reales. Estos residuos representan la parte del objetivo que el modelo actual no ha capturado.

2. **Segundo árbol y subsecuentes:** El proceso continúa con el entrenamiento de un segundo árbol diseñado para predecir los residuos del «Árbol de decisión 1». Cada árbol adicional, representado como el árbol N, se construye con el propósito de corregir los errores residuales del modelo anterior, específicamente del modelo N-1. Este enfoque iterativo permite la construcción de una serie de árboles, donde cada uno está centrado en mejorar las predicciones del árbol anterior.
3. **Asignación de pesos:** En el proceso de entrenamiento de cada árbol, se asignan pesos a las observaciones en función de la magnitud de los errores residuales. Aquellas observaciones con mayores residuos reciben pesos más elevados, lo que permite que el modelo se centre de manera más intensiva en corregir esos errores particulares. Esta asignación de pesos da lugar a «datos ponderados» para cada árbol de decisión, contribuyendo así a un enfoque más preciso en la mejora del modelo.
4. **Predicción Final:** El modelo final integra los resultados de todos los clasificadores débiles para lograr un único clasificador «fuerte» como un «Ensamblado de predicciones».