UNIVERSIDAD PERUANA UNIÓN

ESCUELA DE POSGRADO

Unidad de Posgrado de Ingeniería y Arquitectura

Modelo machine learning para predicción de deserción estudiantil

Tesis para obtener el Título de Segunda Especialidad Profesional de Ingeniería: Estadística Aplicada para Investigación

Autor:

Walter Yataco Cañari

Johnny Prudencio Jacha Rojas

Asesor:

Mg. Johann Alexis Ospina Galindez

Lima, mayo de 2024

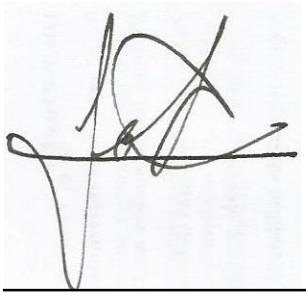
DECLARACIÓN JURADA DE ORIGINALIDAD DE TESIS

Yo (Mag. Johann Alexis Ospina Galindez), docente de la Unidad de Posgrado de Ingeniería y Arquitectura, Escuela de Posgrado de la Universidad Peruana Unión.

DECLARO:

Que la presente investigación titulada: **"Modelo machine learning para predicción de deserción estudiantil"** de los autores Walter Yataco Cañari y Johnny Prudencio Jacha Rojas tiene un índice de similitud de 9 % verificable en el informe del programa Turnitin, y fue realizada en la Universidad Peruana Unión bajo mi dirección.

En tal sentido asumo la responsabilidad que corresponde ante cualquier falsedad u omisión de los documentos como de la información aportada, firmo la presente declaración en la ciudad de Lima, a los 31 días del mes de mayo del año 2024



Mag. Johann Alexis Ospina Galindez)

ACTA DE SUSTENTACIÓN DE TESIS

En Lima, Ñaña, Villa Unión a 14 días del mes de mayo del año 2024, siendo las 8:40 horas, se reunieron de forma online sincrónica, bajo la dirección del presidente del jurado Dra. Ethel Altez Ortiz, secretario Dr. Josué Edison Turpo Chaparro; los demás miembros: PhD. Javier Linkolk López Gonzales, Mg. Lizeth Geanina Huanca Lopez y el asesor Mg. Johann Alexis Ospina Galindez, con el propósito de administrar el acto académico de sustentación de Tesis de la Segunda Especialidad titulada "Modelo machine learning para predicción de deserción estudiantil", conducente a la obtención del Título de Segunda Especialidad Profesional de Ingeniería: Estadística Aplicada para Investigación.

El presidente inició el acto académico de sustentación invitando a los candidatos a hacer uso del tiempo determinado para su exposición. Concluida la exposición, el presidente invitó a los demás miembros del jurado a efectuar las preguntas, cuestionamientos y aclaraciones pertinentes, los cuales fueron absueltos por los candidatos. Luego, se produjo un receso para las deliberaciones y la emisión del dictaminen del jurado. Posteriormente, el jurado procedió a dejar constancia escrita sobre la evaluación en la presente acta, con el dictamen siguiente:

Candidatos: Walter Yataco Cañari y Johnny Prudencio Jacha Rojas

Calificación				Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	18	A-	Con nominación muy bueno	Sobresaliente

Finalmente, el presidente del jurado invitó a los candidatos a ponerse de pie, para recibir la evaluación final. Además, el presidente concluyó el acto académico de sustentación, procediéndose a registrar a registrar las firmas respectivas.

Presidente

Secretario

Asesor

Miembro

Miembro

Candidato

Candidato

INDICE

Modelo machine learning para predicción de deserción estudiantil

Walter Yataco Cañari –
Johnny Prudencio Jacha Rojas

Abstract:

The persistent issue of student dropout negatively impacts the educational sector and society at large. This study presents a machine learning model that leverages data from the National Household Survey to predict student dropout in Peru, integrating a wide range of socio-demographic variables. The research fills a gap in existing literature by providing a model that incorporates socio-demographic variables, an area not fully explored in previous studies. The predictive model aims to identify factors associated with student dropout, aiding educational stakeholders in implementing effective interventions. The findings underscore the model's potential to enhance educational outcomes by enabling early identification of at-risk students, thereby facilitating targeted support. This work contributes to refining predictive models of university dropout rates and suggests the use of ensemble methods to improve the accuracy of single-model predictions. Future research could further explore computational methodologies like deep learning and hybrid models to predict dropout rates and their comparison with this study's outcomes, considering additional influential factors not covered in this research.

Keywords:

Student Dropout, Machine Learning, Predictive Modeling, Socio-Demographic Factors, National Household Survey, Educational Outcomes, Ensemble Methods, Hybrid Models.

1. Introducción

La deserción estudiantil es un problema que impacta negativamente en el sector educativo con efecto negativo en la sociedad. Lograr culminar los estudios ha demostrado que es una oportunidad que impacta en las ofertas laborales y a su vez influye en sus ingresos futuros que aportarán de forma directa en su calidad de vida. **Verástegui (2016)**, menciona que la deserción que los estudiantes rurales tiene factores como el trabajo remunerado, distancia, infraestructura son factores influyentes en la deserción estudiantil.

Los estudiantes desertores han tomado esa decisión por diversos motivos, en algunos casos son decisiones tomadas directamente en otros indirectamente, los factores más influyentes económicos, salud, traslado de ubicación de vivienda, dificultad en recibir conocimientos, social y seguimiento de los padres o apoderado en su educación. **El Peruano (2013)**, informa que el MINEDU apoyara a los estudiantes de 3er, 4to y 5to de segundaría de la Selva de Perú con un bono económico para evitar la deserción estudiantil debido a que realizan trabajos remunerados.

En el año 2020, en Perú se declaró estado de emergencia por la pandemia COVID-19, tras experiencias de países que obtuvieron la pandemia y por estudios e información de especialistas se determinó que el tiempo de aislamientos social sería un plazo no corto, por ello el Ministerio de Educación (MINEDU) analizo la ventajas y desventajas de una educación virtual; inicialmente revisando su soporte informático y presupuesto destinado para una educación virtual, para algunos alumnos este método de enseñanza al ser nuevo y no totalmente experimentando comenzó con dificultades, por el lado Docente reformular las estrategias de enseñanza, en la práctica conllevo a mejoras continuas en su desarrollo; impactando en algunos estudiantes que no lograron continuar sus estudios por diversos factores, uno de ellos el factor informático que corresponde al dispositivo, conectividad, tener un espacio adecuado y otros**. INEI (2022)**, menciona que la deserción estudiantil tuvo un aumento en los años 2020 y 2021 debido a la pandemia COVID-19, fue de 1,4% indicando que solo 110 138 estudiantes de los que estuvieron matriculados en el 2020 (8 088 233 estudiantes) no se matricularon el 2021.

CARE Perú (2023), menciona que el Perú ha obtenido una tasa de deserción escolar del 6.3%, según cifras del Ministerio de Educación (Minedu). Según una encuesta del 2021 del INEI, 22 de cada 100 jóvenes entre 17 y 18 años, no han logrado concluir su educación secundaria, mientras 5 de cada 100 jóvenes entre 13 y 19 años no la ha culminado. Esto significa que existe un gran porcentaje de la población que carece de las habilidades necesarias para tener competitividad en el mercado laboral y contribuir a la productividad del país.

La deserción estudiantil constituye un desafío crítico a nivel mundial, impactando negativamente tanto en las instituciones educativas como en el tejido socioeconómico de comunidades y naciones. Este fenómeno, que trasciende las barreras físicas de las aulas, ejerce una influencia significativa en el desarrollo integral de la sociedad. Ante esta problemática, los modelos de Machine Learning emergen como una solución innovadora

y prometedora, ofreciendo la capacidad de anticipar el riesgo de abandono escolar en los estudiantes.

La eficacia de los modelos de aprendizaje automático reside en su habilidad para procesar y analizar grandes volúmenes de datos, identificando patrones sutiles que frecuentemente son inaccesibles mediante métodos convencionales. Investigaciones llevadas a cabo por **García y Weiss (2020)** enfatizan la importancia de una detección temprana en la implementación de intervenciones personalizadas, un enfoque que es respaldado por **Zhao et al. (2023)**, quien destaca la necesidad de adaptar las estrategias a las necesidades individuales de cada estudiante.

El presente estudio se centra en el desarrollo y validación de un modelo de machine learning robusto y confiable para predecir la deserción estudiantil. Este modelo, que aprovecha los datos de la Encuesta Nacional de Hogares, incorpora un amplio rango de variables sociodemográficas, abordando una laguna significativa en la literatura existente. El estado actual de la investigación se ha enfocado en modelos predictivos como árboles de decisión, bosques aleatorios y modelos de ensamble para perfilar la deserción estudiantil **(Dekker et al., 2009; Flores-Caballero et al., 2020).** No obstante, la integración completa de variables sociodemográficas dentro de los modelos de Machine Learning aún representa un área poco explorada, ofreciendo una oportunidad significativa para mejorar la precisión y la aplicabilidad de estos modelos en contextos reales. Con esta investigación, se busca desarrollar un modelo predictivo basado en Machine Ñearning para identificar los factores asociados a la deserción estudiantil en Perú, utilizando los datos de la Encuesta Nacional de Hogares.

2. Trabajo relacionado

La integración de la inteligencia artificial (IA) en diversos sectores ha sido transformadora, y el ámbito de la educación no es una excepción. La IA y los algoritmos de aprendizaje automático (ML) tienen el potencial de mejorar significativamente el proceso de aprendizaje ofreciendo sistemas predictivos para ayudar a los estudiantes a planificar su trayectoria académica y mejorar su rendimiento **(Zawacki-Richter et al., 2019)**. El uso de la IA en la educación no solo se limita a los entornos académicos tradicionales, sino que también se extiende a la educación especial, donde tiene el potencial de provocar cambios sustanciales en las prácticas de enseñanza (Marino et al., 2023). Además, la aplicación de la IA en la educación no se centra únicamente en las instituciones académicas, sino que también abarca el desarrollo de programas educativos, como los diseñados para la educación media, lo que refleja el impacto multifacético de la IA en el campo de la educación (Park y Kwon, 2023).

El potencial de la IA en la educación se ve reforzado por su impacto en los procesos de toma de decisiones, ya que ofrece la capacidad de proporcionar experiencias de aprendizaje personalizadas y mejorar la calidad general de la educación (Banerjee et al., 2021). Además, el uso de la IA en la educación no se limita al aula, sino que se extiende a la educación clínica, donde se utiliza cada vez más, lo que indica las diversas

aplicaciones de la IA en diferentes ámbitos educativos (Dahmash et al., 2020). Además, la influencia de la IA en la educación no se limita únicamente al proceso de aprendizaje, sino que también se extiende a las percepciones y elecciones profesionales de los estudiantes, como demuestra su impacto en las preferencias de los estudiantes de medicina por la radiología como futura carrera (Lainjo y Tsmouche, 2023).

La predicción del éxito o fracaso en el aprendizaje de un estudiante es uno de los temas más investigados en las disciplinas de minería de datos educativos (EDM) y análisis del aprendizaje (LA). La EDM se centra en el análisis de datos relacionados con el estudio para comprender el comportamiento de los estudiantes, con el fin de proporcionar entornos de aprendizaje más eficaces al revelar información útil para modificar la estructura del curso o para ayudar en la predicción del rendimiento y el comportamiento de los estudiantes (Prenkaj et al., 2020). Por otro lado, LA se ocupa de la medición, recopilación, análisis y presentación de informes de datos y antecedentes de los estudiantes para comprender y mejorar el aprendizaje y los entornos en los que se produce (Siemens & Baker, 2012).

Los métodos de EDM y LA suelen ser el núcleo de los enfoques de predicción actuales en el ámbito educativo. La predicción de la probabilidad de que los estudiantes completen o suspendan un curso, especialmente en las primeras semanas, ha sido uno de los temas de investigación más candentes en la analítica del aprendizaje, al igual que en la minería de datos educativos (Baker & Inventado, 2014). Una vez que se dispone de una predicción fiable del rendimiento, esta puede utilizarse para identificar a los estudiantes débiles y proporcionar retroalimentación, así como para predecir el fracaso de los estudiantes (Prenkaj et al., 2020).

3. Descripción del conjunto de datos y pre procesamiento

Se utilizaron los datos de la Encuesta Nacional de Hogares 2022 condiciones de vida y pobreza (ENAHO.01A) que contiene la información de educación, salud, empleo e ingreso.

En la etapa preliminar de nuestra investigación, se llevó a cabo un meticuloso escrutinio de 511 variables, con el objetivo de determinar su idoneidad para el análisis propuesto. Este proceso implicó un diagnóstico detallado para identificar aquellas variables que cumplían con los criterios de selección establecidos. Una fase importante de este procedimiento fue la evaluación exhaustiva de la integridad de los datos, efectuada durante la etapa de limpieza de los mismos. De acuerdo a esto, se decidió focalizar el estudio en un conjunto de 11 compuesto por 10 variables explicativas, denominadas como 'X', y una variable dependiente, etiquetada como 'Y', que se empleó para representar la variable de interés, en este caso, el fenómeno de 'DESERTOR'. ver Tabla 1.

Tabla 1. Operacionalización de las características.

Característica	Descripción
P303	¿El año pasado estuvo matriculado en algún centro o programa de educación superior?
P301A	¿Cuál es el último año o grado de estudios y nivel de que aprobó?
P305	¿El resultado que obtuvo el año pasado fue?
P306	Este año, ¿Está matriculado en algún centro o programa de educación básica o superior?
P308B	¿Cuál es el año o grado de estudios en el que está matriculado?
P313	¿Cuál es la principal razón por la que no está matriculado o no asiste a algún centro o programa de educación básica o superior?
P314B-1	En el mes anterior ¿Usó usted el servicio de internet?
P314B-2	En el mes anterior, ¿El servicio de internet lo usó a través de una/un?
P314D	¿Usted usa internet al menos?
P316A1	¿En el mes anterior, usted utilizó?
P316B	En los últimos 3 meses, ¿Ha utilizado una computadora, laptop, Tablet o similar?

4. Metodología

Para abordar el desafío de clasificación supervisada propuesto en la creación de un modelo predictivo que identifique la deserción escolar, se han seleccionado y aplicado diversos algoritmos de aprendizaje automático.

4.1.Random Forest Ranfom Forest

Es un algoritmo ampliamente utilizado en el aprendizaje supervisado para tareas de clasificación y regresión. Se basa en la combinación de múltiples árboles de decisión, donde cada árbol se entrena con una porción aleatoria de los datos de entrenamiento originales, lo que contribuye a la robustez del modelo (Breiman, 2001). La biblioteca Scikit-Learn (Sklearn) proporciona la herramienta "RandomForestClassifier", que construye un número de árboles de decisión sobre distintas submuestras del conjunto de datos y mejora la precisión de las predicciones finales a través del promedio, a la vez que previene el riesgo de sobreajuste**(Cutler et al., 2007)**. Este metaestimador ofrece diversos parámetros ajustables para optimizar su eficacia **(Hartini et al., 2021)**.

4.2.Bagging

Es un algoritmo conjunto (ensemble) de Machine Learning que combina las predicciones de distintos clasificadores. Aquí se ajustan múltiples modelos, cada uno con un subconjunto distinto de los datos de entrenamiento (Breiman, 1996). Para realizar la predicción, todos los modelos participan aportando su predicción. Como valor final, se toma la media de todas las predicciones (variables continuas) o la clase más frecuente (variables categóricas).

4.3.XGBoost

Es una técnica avanzada de aprendizaje supervisado que ha demostrado ser efectiva tanto en problemas de clasificación como de regresión. Se trata de una versión de código abierto de los árboles de decisión mejorados mediante gradientes, que optimiza la función de pérdida a través del descenso por gradiente para mejorar la precisión (Chen, 2016). Este algoritmo, conocido por su eficiencia y velocidad, requiere la instalación del paquete xgboost en el sistema o entorno de desarrollo que se esté utilizando, disponible en su documentación oficial (Chen, 2016). Utilizando el modelo "XGBClassifier", este algoritmo ofrece la posibilidad de ajustar diversos hiperparámetros críticos para optimizar el rendimiento y la eficacia de las predicciones (Chen, 2016).

4.4.Medidas de desempeño

Con el propósito de evaluar el rendimiento del modelo predictivo de deserción estudiantil para los algoritmos entrenados de clasificación binaria en el aprendizaje supervisado, es necesario evaluarlos en función de distintas métricas de desempeño. En este caso de estudio, la métrica principal implementada de Machine Learning es la matriz de confusión “Confusion Matrix” de la librería de Sklearn, la cual permite evaluar la precisión de la predicción en el proceso de clasificación binaria (1 - desertores y 0 - no desertores) usando una tabla cruzada, arrojando información relevante como verdaderos positivos (True positive - TP), falsos positivos (False positive - FP), verdaderos negativos (True negative - TN) y falsos negativos (False negative - FN). Esta es usada debido a la gran importancia

de visualizar el rendimiento en la predicción de los verdaderos positivos y falsos negativos.

Adicionalmente, se realiza el cálculo del puntaje de clasificación de precisión "Accuracy" como la proporción de predicciones correctas sobre el número total de las predicciones realizadas por el modelo con respecto a la matriz de confusión. La sensibilidad "Recall" o tasa de verdaderos positivos, como la proporción de aquellos predichos como positivos (en este caso a los posibles desertores) entre los verdaderos positivos con respecto a la matriz de confusión. La puntuación promedia ponderada de la precisión y recuperación "F1 Score" la cual corresponde al cálculo de la media armónica de la precisión y la sensibilidad y el "AUC" correspondiente al área bajo la curva característica de operativa (ROC). Los valores de AUC oscilan entre 0,5 y 1,0. Cuando el valor de AUC se aproxima a 1, indica que el modelo tiene un mejor rendimiento, mientras que un valor inferior a 0,5 indica un rendimiento deficiente y la inclinación del ROC debe ser alta.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Todas las métricas descritas anteriormente, fueron implementadas para complementar la validación a partir de datos de prueba, en cada uno de los modelos utilizados de Machine Learning.

5. Resultados

La Tabla 1 muestra el análisis de la población estudiada, compuesta por 2953 encuestados, se observa un compromiso total con la educación superior, evidenciado por la matriculación del 100% de los participantes en programas o centros de educación superior el año anterior. Este interés se refleja en el nivel educativo alcanzado, donde una mayoría significativa, el 83.47%, aún no ha completado sus estudios universitarios, sugiriendo que la mayoría son estudiantes activos en este nivel.

A pesar de los desafíos, los resultados académicos del año anterior fueron predominantemente positivos, con un 93.12% de aprobación. Sin embargo, este año, solo el 77.68% de los encuestados continúa matriculado en algún nivel de educación, ya sea básica o superior, marcando una disminución en comparación con el año anterior. La distribución actual de niveles educativos es diversa, abarcando desde la educación inicial hasta estudios de posgrado, aunque con una concentración notable en la educación superior, tanto universitaria como no universitaria.

El uso de internet es casi universal entre los encuestados, con una mayoría accediendo a este servicio tanto desde el hogar como en el trabajo, lo que subraya la importancia de la conectividad digital en sus vidas. La preferencia por dispositivos como computadoras y laptops para acceder a internet es clara, aunque el uso diario de internet casi alcanza la totalidad de la muestra, demostrando su integración en las rutinas diarias.

El teléfono celular emerge como un dispositivo personal esencial, con casi todos los participantes usando el suyo para diversas actividades. Adicionalmente, el uso reciente de tecnología, especificado por el 91.39% que ha utilizado computadoras, laptops, tablets o similares en los últimos tres meses, refleja una familiaridad y dependencia considerable hacia la tecnología digital dentro de esta población.

Este panorama destaca no solo el acceso generalizado y la integración de la tecnología en la educación y la vida cotidiana de los encuestados, sino también la persistente aspiración y el progreso hacia la culminación de la educación superior, delineando un perfil de una comunidad estudiantil que, a pesar de los retos académicos y la diversidad en los niveles de estudios, muestra un compromiso continuo con su desarrollo educativo y una adaptación proactiva a las herramientas digitales como medio indispensable para su avance y conexión con el mundo.

Tabla 2. Distribución de las características.

Característica	n	%
¿El año pasado estuvo matriculado en algún centro o programa de educación superior?		
Si	2953	100
No	0	0
¿Cuál es el último año o grado de estudios y nivel de que aprobó?		
Secundaria completa	76	2.57
Sup. No universitaria completa	8	0.27
Sup. Universitaria incompleta	2465	83.47
Sup. Niversitaria completa	395	13.37
Maestría/Doctorado	9	0.3
¿El resultado que obtuvo el año pasado fue?		
¿Aprobado?	2750	93.12
¿Desaprobado?	7	0.23
¿Retirado?	27	0.91
¿Otro?	169	5.72
¿No aprueba, ni desaprueba (inicial)?	0	0
Este año, ¿Está matriculado en algún centro o programa de educación básica o superior?		
Si	2294	77.68
No	659	22.32
¿Cuál es el año o grado de estudios en el que está matriculado?		
Educación inicial	88	2.98
Primaria	552	18.69
Secundaria	529	17.91
Básica especial	16	0.54
Sup. No universitaria	532	18.01
Sup. Universitaria	522	17.67
Maestría/Doctorado	55	1.86
Sin dato	659	22.31
En el mes anterior ¿Usó usted el servicio de internet?		
El hogar	1241	42.02
El trabajo	1693	57.33
Sin dato	19	0.64
En el mes anterior, ¿El servicio de internet lo usó a través de una/un?		
Computadora	2718	82.04
Laptop	216	7.31
Sin dato	19	0.64
¿Usted usa internet al menos?		
Una vez al día	2890	97.86
Una vez a la semana	44	1.5
Sin dato	19	0.64
¿En el mes anterior, usted utilizó?		
Teléfono celular propio	2888	97.79
Teléfono celular de un familiar o amigo/a	65	2.2

En los últimos 3 meses, ¿Ha utilizado una computadora, laptop, Tablet o similar?		
Si	2699	91.39
No	254	8.6

5.1.Modelación

El estudio utilizó el lenguaje de programación Python (versión 3.12.1), empleando específicamente el paquete scikit-learn (versión 1.5) para desarrollar una técnica de apilamiento que combina eficazmente los algoritmos RF, XGBoost, GB y FNN para predecir las tasas de abandono universitario (Probst et al., 2019). Para mitigar el sobreajuste, los investigadores utilizaron la validación cruzada durante el proceso de entrenamiento del modelo (Gaïffas et al., 2021). Los datos se dividieron en un 80% para el entrenamiento y un 20% para las pruebas, siguiendo la recomendación de Joseph & Vakayil (Athey et al., 2019). Además, se implementó una estrategia de validación cruzada de diez veces para mejorar la fiabilidad de los resultados y combatir el sobreajuste (Athey et al., 2019).

5.2.Estimación de hiperparamétros

El enfoque adoptado en este estudio consiste en un modelo compuesto que requiere la calibración de varios parámetros clave. Para ello, se llevó a cabo una optimización meticulosa mediante técnicas de búsqueda en cuadrícula, con el objetivo de afinar los parámetros esenciales y, así, elevar la precisión de las predicciones **(Feurer & Hutter, 2019)**. Este modelo incorpora en su primera capa algoritmos de aprendizaje ensambl como RF y XGBoost, seleccionados por su eficacia para generar resultados más precisos que los que podría ofrecer un predictor individual **(Li et al., 2021)**.

5.3.Comparación del rendimiento de los modelos

Para evaluar la eficacia de los modelos implementados. Los resultados se comparan a partir de los resultados de entrenamiento y prueba producidos por RF, GB, XGBoost y Stacking ensemble. En la evaluación de varios modelos de aprendizaje automático, se han medido tanto el rendimiento en un conjunto de pruebas como en uno de validación usando la precisión (accuracy) como la métrica de evaluación. El modelo KNeighborsDist ha mostrado el mejor rendimiento en el conjunto de pruebas con un score de 0.8646 y en el conjunto de validación con un score de 0.8710. Este modelo ha demostrado ser eficiente en términos de tiempo de predicción y tiempo de ajuste, lo que indica un balance entre precisión y eficiencia computacional.

El modelo KNeighborsUnif, aunque ligeramente inferior en el conjunto de pruebas con un score de 0.8629, ha mostrado una ligera mejora en el conjunto de validación con un score de 0.8731. Esto podría sugerir que el modelo KNeighborsUnif generaliza mejor a datos no vistos anteriormente.

El WeightedEnsemble_L2, a pesar de tener el mismo rendimiento que KNeighborsUnif en términos de precisión, tiene un tiempo de ajuste considerablemente más alto, lo que podría ser un factor limitante si el tiempo de entrenamiento es una consideración crítica.

El NeuralNetFastAI, aunque comparable en precisión en los conjuntos de prueba y validación, tiene el mayor tiempo de ajuste de todos los modelos evaluados, lo que podría ser una desventaja en términos de eficiencia, especialmente en un entorno de producción donde el tiempo y los recursos son críticos.

Los modelos basados en gradient boosting como LightGBMXT, CatBoost y XGBoost también han demostrado una alta precisión en el conjunto de validación, lo que los hace candidatos adecuados para problemas en los que se requiere un equilibrio entre rendimiento y velocidad de predicción.

Por último, los modelos de ensamble como RandomForest y ExtraTrees (tanto Gini como Entr) han mostrado una precisión ligeramente más baja en comparación con los otros modelos. Estos modelos también han incurrido en los tiempos de predicción y ajuste más largos, lo que podría no ser ideal en situaciones en las que el tiempo de respuesta es crítico.

Tabla 3. Resultados de la precisión de los métodos implementados.

Modelo	Score Test	Score Val	Métrica Evaluación	Tiempo Pred. Test	Tiempo Pred. Val	Tiempo Ajuste
KNeighborsDist	0.8646	0.8710	accuracy	0.0192	0.0241	0.0141
KNeighborsUnif	0.8629	0.8731	accuracy	0.0171	0.0213	0.0111
WeightedEnsemble_L2	0.8629	0.8731	accuracy	0.0201	0.0231	1.1736
NeuralNetFastAI	0.8613	0.8710	accuracy	0.0430	0.0148	2.6550
LightGBMXT	0.8596	0.8731	accuracy	0.0070	0.0063	0.6242
CatBoost	0.8579	0.8731	accuracy	0.0049	0.0177	1.4303
XGBoost	0.8579	0.8710	accuracy	0.0304	0.0059	0.2571
LightGBMLarge	0.8562	0.8710	accuracy	0.0064	0.0040	0.4886
NeuralNetTorch	0.8562	0.8710	accuracy	0.0236	0.0124	3.5972
LightGBM	0.8545	0.8647	accuracy	0.0040	0.0072	0.6287
RandomForestGini	0.8528	0.8584	accuracy	0.1420	0.1604	1.5014
ExtraTreesGini	0.8528	0.8562	accuracy	0.1528	0.1210	1.9857
RandomForestEntr	0.8528	0.8584	accuracy	0.1617	0.1378	1.2517
ExtraTreesEntr	0.8528	0.8562	accuracy	0.1714	0.1205	0.8421

En la **Tabla 4** se muestran las métricas de rendimiento del mejor modelo obtenido (KNeighborsDist). Se destaca por su eficiencia en clasificación, demostrando alta precisión con un valor de 0.8646, lo que refleja su capacidad para realizar predicciones correctas en una gran mayoría de los casos.

La métrica de precisión balanceada de 0.8231 sugiere que este rendimiento es consistente a través de clases desiguales, una cualidad importante para conjuntos de datos con

distribuciones de clases desbalanceadas. Además, un ROC AUC de 0.8205 indica una capacidad confiable del modelo para diferenciar entre las clases positivas y negativas.

El modelo también logra un excelente equilibrio entre la precisión y el Recall, con un score F1 de 0.9121, lo que demuestra que es particularmente efectivo para identificar los casos positivos reales mientras minimiza los falsos positivos, como lo evidencia su alta precisión de 0.9284 y un recall impresionante de 0.8963.

Estos resultados indican que el KNeighborsDist es un modelo robusto y confiable para aplicaciones de clasificación, proporcionando garantías tanto en la identificación correcta de la clase de interés como en la generalización efectiva a datos no vistos previamente.

Tabla 4. Métricas del mejor modelos (KNeighborsDist).

Métrica	Valor
Accuracy	0.8646
Balanced_accuracy	0.8231
ROC AUC	0.8205
F1 score	0.9121
Precision	0.9284
Recall	0.8963

El análisis de las métricas de rendimiento revela que los modelos de clasificación binaria escogidos son adecuados para pronosticar la permanencia o deserción estudiantil en cursos específicos, aun en presencia de un conjunto de datos restringido y con características de entrada limitadas. No obstante, para afirmar la capacidad predictiva de los clasificadores en relación con la deserción estudiantil, es imperativo explorar un espectro más amplio de indicadores de rendimiento. Esta perspectiva se alinea con las deducciones de investigaciones previas en las esferas de inteligencia artificial y aprendizaje automático.

Sin embargo, el caso de estudio abordado y sus deducciones enfrentan ciertas restricciones. Una de ellas, ya señalada, es la escasez del volumen de datos. En contraste con otros campos de aplicación de modelos de aprendizaje automático, el incremento del volumen de datos en el sector educativo no es sencillo mediante la integración de fuentes diversas. Esto se debe a que los registros individuales deben reflejar con precisión los logros de aprendizaje o comportamientos estudiantiles, lo que conlleva a conjuntos de datos más reducidos que los preferidos por los algoritmos de aprendizaje automático.

A menudo, esta barrera se supera solo con una meticulosa planificación de la investigación, similar a la adoptada en estudios convencionales de tecnología educativa, donde se asegura o estima un volumen de datos adecuado previo al experimento.

5.4. Importancia de características

La **tabla 5** muestra los resultados de la importancia de características del mejor modelo obtenido. la característica "¿Cuál es el año o grado de estudios en el que está matriculado?" destaca con una importancia significativa (0.2179) y una muy baja desviación estándar, acompañada de un valor-p casi nulo, lo que indica una influencia estadísticamente significativa sobre la variable de interés. Esto sugiere que el nivel de estudios actual tiene un fuerte impacto predictivo, potencialmente reflejando la relevancia de la etapa educativa en el fenómeno de estudio.

Por otro lado, "¿Cuál es el último año o grado de estudios y nivel de que aprobó?" también muestra relevancia, aunque en menor medida (0.0034), con un valor-p bajo, lo que indica que los logros educativos previos también contribuyen significativamente a las predicciones del modelo. Las demás características relacionadas con el uso de internet y tecnología (como computadoras o laptops) tienen importancias cercanas a cero y valores-p de 0.5, lo que sugiere que no hay evidencia estadística que respalde su relevancia predictiva.

La característica "¿El resultado que obtuvo el año pasado fue?" tiene una importancia ligeramente negativa, lo que podría sugerir una relación inversa con la variable objetivo, aunque su alto valor-p sugiere que este resultado no es estadísticamente significativo. Este análisis refleja la importancia crítica de la trayectoria académica sobre otros factores tecnológicos en el ámbito educativo, al menos dentro del conjunto de datos analizado.

Tabla 5. Importancias de Características.

Característica	Importancia	Desviación Estándar	Valor-p	I.C95%	
¿Cuál es el año o grado de estudios en el que está matriculado?	0.2179357	0.010861	0.000000737	0.240298	0.195573
¿Cuál es el último año o grado de estudios y nivel de que aprobó?	0.003384095	0.001692	0.005528247	0.006868	-0.000100
¿Usted usa internet al menos?	2.22e-17	0.001196	0.5	0.002464	-0.002464
Este año, ¿Está matriculado en algún centro o programa de educación básica o superior?	0.0	0.0	0.5	0.0	0.0
En el mes anterior ¿Usó usted el servicio de internet?	0.0	0.0	0.5	0.0	0.0
En el mes anterior, ¿El servicio de internet lo usó a través de una/un?	0.0	0.0	0.5	0.0	0.0
¿En el mes anterior, usted utilizó?	0.0	0.0	0.5	0.0	0.0

En los últimos 3 meses, ¿Ha utilizado una computadora, laptop, Tablet o similar?	0.0	0.0	0.5	0.0	0.0
¿El resultado que obtuvo el año pasado fue?	-0.0003384095	0.000757	0.8130495	0.001220	-0.001896

Conclusiones

La deserción educativa representa un desafío crítico y adverso dentro del entorno educativo, afectando negativamente el proceso educativo. Esta problemática ha captado la atención de diversos actores del ecosistema educativo, incluyendo familias, entidades gubernamentales y otros interesados, debido a sus profundas repercusiones. A pesar de los esfuerzos implementados para mitigar esta situación, las consecuencias de la deserción continúan presentándose. La capacidad de pronosticar con exactitud la deserción podría contribuir significativamente a reducir su impacto social y económico. Las herramientas analíticas basadas en aprendizaje automático tienen el potencial de identificar y prever con eficacia los factores determinantes de la deserción, tales como el progreso académico del estudiante, las condiciones de aprendizaje, las variables demográficas y sociales, y otros aspectos relacionados con el soporte educativo, ofreciendo así valiosa inteligencia para comprender y prevenir la deserción escolar.

La capacidad de prever con precisión el desempeño académico es clave para que los estudiantes se mantengan enfocados en sus objetivos educativos, minimizando la posibilidad de deserción escolar. Tal anticipación permite a los administradores educativos utilizar los datos proyectados para tomar decisiones fundamentadas respecto a la promoción estudiantil o la consideración de medidas alternativas para aquellos con desempeño insuficiente, como ofrecer oportunidades de recuperación o exámenes de segunda oportunidad. Una detección temprana de los patrones de comportamiento que podrían resultar en una deserción permite a los educadores intervenir de manera temprana y adoptar estrategias preventivas. Este trabajo aporta a la optimización de los modelos predictivos del abandono universitario, posibilitando una evaluación más precisa del riesgo de deserción y contribuyendo a la mitigación de sus consecuencias en el rendimiento estudiantil. Asimismo, esta investigación avanza en la implementación de métodos de ensamblaje de modelos para superar las limitaciones de las predicciones basadas en un solo modelo, mejorando así la exactitud en la identificación de estudiantes en riesgo de abandono escolar.

La aplicación de la metodología propuesta en este estudio ofrece la posibilidad de disminuir la incidencia de deserción estudiantil al identificar precozmente a los estudiantes en riesgo y los elementos críticos que contribuyen a su situación. Una vez reconocidos estos estudiantes, los profesionales de la educación pueden unir esfuerzos

para implementar soluciones efectivas que aborden directamente la raíz del problema. Asimismo, se puede poner en práctica una variedad de tácticas para fomentar la motivación y el rendimiento estudiantil, lo cual podría facilitar la culminación exitosa de sus programas académicos. Investigaciones futuras podrían explorar metodologías computacionales avanzadas, como el aprendizaje automático profundo y modelos híbridos, para prever la deserción y evaluar su eficacia en comparación con los hallazgos presentados aquí. Es crucial considerar la influencia de factores adicionales que este estudio no abarcó, recomendándose una investigación exhaustiva sobre la selección de características que pueda respaldar a los educadores en la gestión eficiente de la deserción escolar.

Referencias

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. The Annals of Statistics, 47(2). https://doi.org/10.1214/18-aos1709

Banerjee, M., Chiew, D., Patel, K. T., Johns, I., Chappell, D., Linton, N., … & Zaman, S. (2021). The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in london (uk) and recommendations for trainers. BMC Medical Education, 21(1). https://doi.org/10.1186/s12909-021-02870-x

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. *Learning Analytics: From Research to Practice*, 61–75. https://doi.org/10.1007/978-1-4614-3305-7_4

Breiman, L. (2001). Untitled. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/a:1010933404324

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140. https://doi.org/10.1007/bf00058655

CARE Perú. (2023). *5 cifras alarmantes de la educación en el Perú : CARE Perú*. https://care.org.pe/5-cifras-alarmantes-de-la-educacion-en-el-peru/

Chen, T. and Guestrin, C. (2016). Xgboost: a scalable tree boosting system.. https://doi.org/10.48550/arxiv.1603.02754

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., … & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792. https://doi.org/10.1890/07-0539.1

Dahmash, A. B., Alabdulkareem, M., Alfutais, A., Kamel, A., Alkholaiwi, F., Al-Shehri, S., … & Almoaiqel, M. (2020). Artificial intelligence in radiology: does it impact medical students preference for radiology as their future career?. BJR|Open, 2(1), 20200037. https://doi.org/10.1259/bjro.20200037

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. *EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining*, *May 2014*, 41–50. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b0185d5ffa59106088125d82a3ca296bc5b09c87

EL PERUANO. (2013). *Darán bono a familias cuyos hijos culminen sus estudios*. https://elperuano.pe/noticia/10277-daran-bono-a-familias-cuyos-hijos-culminen-sus-estudios

Feurer, M., & Hutter, F. (2019). *Hyperparameter Optimization*. 3–33. https://doi.org/10.1007/978-3-030-05318-5_1

Flores-Caballero, D., Serna-Alarcón, V., Aliaga-Cajan, J., Sandoval-Ato, R., Benites-Meza, J. K., & Barboza, J. J. (2020). Predictive model of neonatal hypoglycemia in a public hospital north of Peru: Case-control study. *Revista del Cuerpo Medico Hospital Nacional Almanzor Aguinaga Asenjo*, *13*(3), 286–290. https://doi.org/10.35434/rcmhnaaa.2020.133.739

García, E., & Weiss, E. (2020). *La Importancia Del Absentismo Escolar Para El Desarrollo Y El Desempeño Educativos*. https://www.fundacionareces.es/recursos/doc/portal/2018/03/20/la-importancia-del-absentismo-escolar.pdf

Gaïffas, S., Merad, I., & Yu, Y. (2021). Wildwood: a new random forest algorithm.. https://doi.org/10.48550/arxiv.2109.08010

Hartini, S., Rustam, Z., Saragih, G. S., & Vargas, M. J. S. (2021). Estimating probability of banking crises using random forest. *IAES International Journal of Artificial Intelligence*, *10*(2), 407–413. https://doi.org/10.11591/IJAI.V10.I2.PP407-413

INEI, I. nacional de E. e I. (2022). Pandemia y desercion escolar en la educacion basica regular: Factores asociados y posibles efectos, 2017-2021. *Instituto nacional de Estadística e Informática*, 88–1. https://www.inei.gob.pe/media/MenuRecursivo/investigaciones/desercion-escolar.pdf

Marino, M. T., Vasquez, E., Dieker, L., Basham, J. D., & Blackorby, J. (2023). The future of artificial intelligence in special education technology. Journal of Special Education Technology, 38(3), 404-416. https://doi.org/10.1177/01626434231165977

Lainjo, B. and Tsmouche, H. (2023). Impact of artificial intelligence on higher learning institutions. International Journal of Education, Teaching, and Social Sciences, 3(2), 96-113. https://doi.org/10.47747/ijets.v3i2.1028

Li, C., Zhou, L., & Xu, W. (2021). Estimating aboveground biomass using sentinel-2 msi data and ensemble algorithms for grassland in the shengjin lake wetland, China. *Remote Sensing*, *13*(8). https://doi.org/10.3390/rs13081595

Park, W. and Kwon, H. (2023). Implementing artificial intelligence education for middle school technology education in republic of korea. International Journal of Technology and Design Education. https://doi.org/10.1007/s10798-023-09812-2

Plataforma Nacional de Datos Abiertos. (2022). *Encuesta Nacional de Hogares (ENAHO) 2022 - [Instituto Nacional de Estadística e Informática – INEI] | Plataforma Nacional de Datos Abiertos*. https://www.datosabiertos.gob.pe/dataset/encuesta-nacional-de-hogares-enaho-2022-instituto-nacional-de-estadística-e-informática-–

Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses. *ACM Computing Surveys (CSUR)*, *53*(3). https://doi.org/10.1145/3388792

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. WIREs Data Mining and Knowledge Discovery, 9(3). https://doi.org/10.1002/widm.1301

Siemens, G., & Baker, R. S. J. D. (2012). Learning analytics and educational data mining: Towards communication and collaboration. *ACM International Conference Proceeding Series*, 252–254. https://doi.org/10.1145/2330601.2330661

Verástegui Arteaga, W. J. (2016). Deserción escolar: evolución, causas y relación con la tasa de conclusión de educación básica. *Ministerio de Educación*, 1–2. https://hdl.handle.net/20.500.12799/8772

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators?. International Journal of Educational Technology in Higher Education, 16(1). https://doi.org/10.1186/s41239-019-0171-0

Zhao, K., Chen, N., Liu, G., Lun, Z., & Wang, X. (2023). School climate and left-behind children's achievement motivation: The mediating role of learning adaptability and the moderating role of teacher support. *Frontiers in Psychology*, *14*(January), 1–13. https://doi.org/10.3389/fpsyg.2023.1040214