Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Facultad De Ingeniería De Sistemas e Informática

Modelo predictivo y sistema de recomendaciones personalizadas para mejorar la calificación académica en los estudiantes de Ingeniería de Software de la UNMSM mediante aprendizaje automático

Proyecto de Investigación presentado en el curso Desarrollo de Tesis I para la carrera profesional de Ingeniería de Software

Autor
Roddy David Pérez Acosta

Profesor
Dr. José Alfredo Herrera Quispe

Lima, Julio, 2025

PERÚ

ÍNDICE

Resumen

El presente trabajo propone un modelo predictivo y un sistema de recomendaciones personalizadas orientado a mejorar la calificación académica de estudiantes de Ingeniería de Software de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos (UNMSM). Se utilizaron algoritmos de aprendizaje automático supervisado, específicamente Random Forest y Gradient Boosting, para predecir la nota semestral de un estudiante a partir de características psicoeducativas, socioeconómicas y conductuales extraídas de un conjunto de datos reales.

El modelo se integró en una plataforma interactiva desarrollada con Streamlit, que no solo permite realizar predicciones en tiempo real, sino también generar recomendaciones personalizadas mediante simulaciones. Estas recomendaciones indican cómo ciertas modificaciones en variables como las horas de estudio, el sueño, la asistencia o el uso de redes sociales podrían impactar positivamente en la nota estimada.

Los resultados obtenidos muestran métricas de rendimiento satisfactorias, alcanzando un coeficiente de determinación (R^2) de hasta 0.81 con Gradient Boosting. Este enfoque demuestra el potencial del aprendizaje automático como herramienta de apoyo en la toma de decisiones académicas, ofreciendo información valiosa tanto para estudiantes como para instituciones educativas en la búsqueda de estrategias de mejora del rendimiento académico.

Abstract

This work proposes a predictive model and a personalized recommendation system aimed at improving the academic performance of Software Engineering students from the Faculty of Systems Engineering and Informatics at the National University of San Marcos (UNMSM). Supervised machine learning algorithms, specifically Random Forest and Gradient Boosting, were used to predict students' semester grades based on psychoeducational, socioeconomic, and behavioral features extracted from a real dataset.

The model was integrated into an interactive platform developed using Streamlit, which not only enables real-time predictions but also generates personalized recommendations through simulations. These recommendations indicate how changes in variables such as study hours, sleep, attendance, or social media usage can positively impact the predicted grade.

The results show satisfactory performance metrics, reaching a coefficient of determination (R^2) of up to 0.81 with Gradient Boosting. This approach demonstrates the potential of machine learning as a decision-support tool in academic settings, providing valuable insights for both students and educational institutions in the pursuit of strategies to enhance academic performance.

Capítulo I

Introduccion

1.1 Introducción

En los últimos años, el uso de técnicas de *aprendizaje automático* ha revolucionado diversos sectores, incluida la educación superior. En particular, la predicción del *rendimiento académico* de los estudiantes se ha convertido en un tema de creciente interés dentro de la comunidad científica y académica, ya que permite anticipar resultados académicos desfavorables y, en consecuencia, aplicar medidas preventivas o correctivas oportunas Pan and Dai (2024); Zhu et al. (2025); Alsubhi et al. (2023); Hashim et al. (2020). Este enfoque resulta especialmente valioso en contextos universitarios, donde las tasas de deserción, bajo rendimiento o dificultades en la adaptación curricular representan desafíos persistentes para instituciones educativas de todo el mundo Zhao et al. (2023); Sekeroglu et al. (2021); Albreiki et al. (2021).

Diversas investigaciones han demostrado la eficacia de los modelos de *aprendizaje automático supervisado* para estimar con precisión el rendimiento académico de los es-

tudiantes a partir de datos heterogéneos, que incluyen variables socioeconómicas, psicoeducativas y conductuales Nizar et al. (2024); Oyedeji et al. (2020); Yang et al. (2025); Ahmed (2024). Algoritmos como *Random Forest*, *Gradient Boosting* y *redes neuronales* han mostrado buenos niveles de precisión y generalización en tareas de predicción, superando en muchos casos a métodos estadísticos tradicionales Hashim et al. (2020); Wang and Yu (2025); Buenaño-Fernández et al. (2019). No obstante, la mayoría de estudios se han limitado a realizar predicciones sin generar retroalimentación útil para el estudiante, lo cual reduce el potencial de estos modelos como herramientas activas de mejora académica.

Frente a esta limitación, emergen propuestas que combinan modelos predictivos con sistemas de *recomendaciones personalizadas*, permitiendo no solo anticipar el desempeño de un estudiante, sino también brindarle estrategias concretas para mejorarlo Atalla et al. (2023); Alsubaie (2023); Nachouki et al. (2023). Este enfoque abre nuevas posibilidades en la toma de decisiones pedagógicas basadas en evidencia, especialmente cuando se integran en plataformas interactivas que ofrecen visualizaciones, simulaciones y retroalimentación personalizada en tiempo real.

En este contexto, la presente investigación propone el desarrollo de un modelo predictivo basado en algoritmos de *aprendizaje automático supervisado —Random Forest* y *Gradient Boosting*— para estimar la calificación semestral de estudiantes de Ingeniería de Software de la Universidad Nacional Mayor de San Marcos. Además, se construye un sistema de recomendaciones personalizadas, que mediante simulaciones interactivas, permite visualizar cómo la modificación de ciertas variables —como las horas de estudio, el tiempo de sueño, la asistencia a clases o el uso de redes sociales— puede influir en la nota estimada.

El sistema fue implementado en una plataforma desarrollada con *Streamlit*, que facilita la interacción con el usuario final y la interpretación de los resultados por parte de estudiantes, docentes o autoridades académicas.

Este trabajo tiene como finalidad demostrar el potencial del *aprendizaje automático* no solo como herramienta predictiva, sino también como un instrumento de intervención y mejora continua en el ámbito educativo, promoviendo una educación más personalizada, informada y centrada en el estudiante.

1.2 Planteamiento del problema

1.2.1 Determinación del problema

1.3 Determinación del problema

La presente investigación se enfoca en abordar las limitaciones existentes en los enfoques tradicionales utilizados para estimar el rendimiento académico de los estudiantes universitarios, particularmente en el contexto de la carrera de Ingeniería de Software de la Universidad Nacional Mayor de San Marcos (UNMSM). Si bien los métodos estadísticos convencionales han permitido obtener aproximaciones generales del desempeño estudiantil, presentan dificultades para capturar de manera precisa las relaciones complejas entre múltiples factores psicoeducativos, socioeconómicos y conductuales que influyen directamente en la calificación académica.

El uso de algoritmos de *aprendizaje automático* ha demostrado un potencial significativo en la modelación de estas relaciones no lineales, permitiendo una mayor precisión en la predicción de resultados académicos. No obstante, la mayoría de estudios previos

se centran únicamente en la predicción de las notas, sin proporcionar orientación práctica al estudiante sobre cómo mejorar su rendimiento. Esta limitación reduce el impacto real de los modelos predictivos, al no convertir la información obtenida en estrategias accionables.

Asimismo, en el contexto institucional de la UNMSM, no existen sistemas integrados que permitan no solo predecir la nota semestral de un estudiante a partir de variables como las horas de estudio, el sueño, la asistencia o el uso de redes sociales, sino también generar recomendaciones personalizadas basadas en simulaciones interactivas que indiquen cómo modificar dichas variables para obtener una mejora en la calificación.

Por otro lado, la falta de plataformas accesibles y visuales limita la interacción entre el modelo predictivo y los usuarios finales (estudiantes, docentes, asesores académicos), dificultando la toma de decisiones informadas y basadas en evidencia.

Por tanto, se identifica una necesidad concreta: desarrollar un modelo predictivo confiable, apoyado en algoritmos supervisados como *Random Forest* y *Gradient Boosting*, que no solo anticipe el desempeño académico de los estudiantes, sino que se integre a un sistema interactivo capaz de brindar retroalimentación y recomendaciones personalizadas para mejorar su calificación académica.

1.3.1 Formulación del problema

"Baja capacidad de intervención de los modelos predictivos tradicionales en el rendimiento académico universitario"

En los últimos años, el uso de modelos de *aprendizaje automático* ha cobrado relevancia en el ámbito educativo por su capacidad para predecir con precisión el rendimiento académico de los estudiantes Zhu et al. (2025); Hashim et al. (2020). Algoritmos como

Random Forest y *Gradient Boosting* han demostrado ser herramientas eficaces para estimar calificaciones a partir de variables psicoeducativas, conductuales y socioeconómicas. Sin embargo, la mayoría de estos modelos se enfocan exclusivamente en la predicción, sin ofrecer retroalimentación útil o recomendaciones prácticas al estudiante, lo cual limita su impacto real como herramientas de mejora académica Oyedeji et al. (2020); Ahmed (2024).

Actualmente, en la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos (UNMSM), no existen soluciones tecnológicas que integren modelos predictivos con sistemas interactivos de recomendaciones personalizadas. Esta carencia impide anticipar situaciones de bajo rendimiento o aplicar estrategias de mejora fundamentadas en datos reales.

A pesar del avance de los algoritmos de *machine learning*, el rendimiento académico sigue siendo abordado como una variable aislada, sin tomar en cuenta el potencial de simulación de estos modelos para sugerir acciones concretas de mejora. En ese sentido, la implementación de un sistema que no solo prediga la calificación final del semestre, sino que también evalúe distintos escenarios de modificación de variables clave (como horas de estudio, sueño, uso de redes sociales o asistencia a clases) podría representar un aporte significativo en la personalización del aprendizaje universitario Buenaño-Fernández et al. (2019); Atalla et al. (2023).

La presente investigación se propone cerrar esta brecha mediante el diseño de un modelo predictivo basado en *aprendizaje automático supervisado*, complementado con un sistema de recomendaciones personalizadas, implementado en una plataforma interactiva desarrollada con *Streamlit*. Este enfoque busca no solo anticipar el desempeño académico del estudiante, sino también proporcionarle herramientas prácticas de intervención, con

base en simulaciones personalizadas.

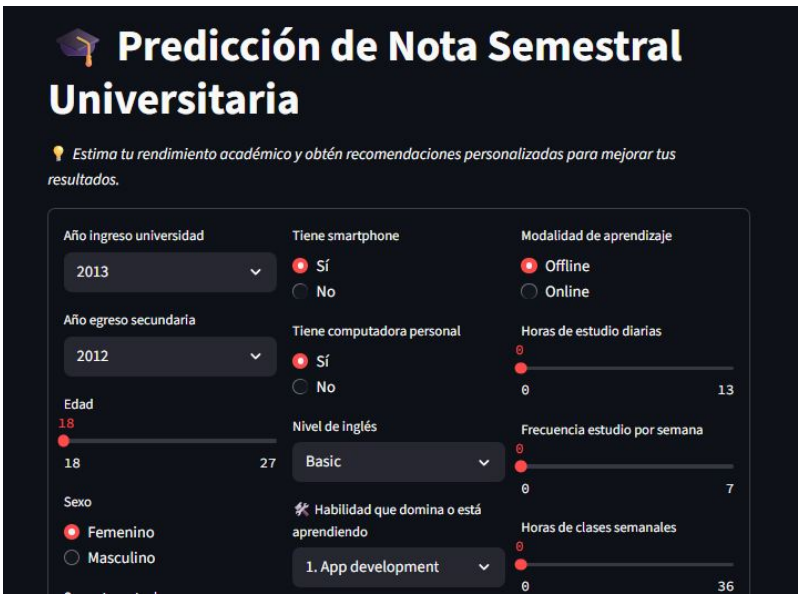

Figura 1.1: Interfaz de la plataforma interactiva para predicción y recomendaciones

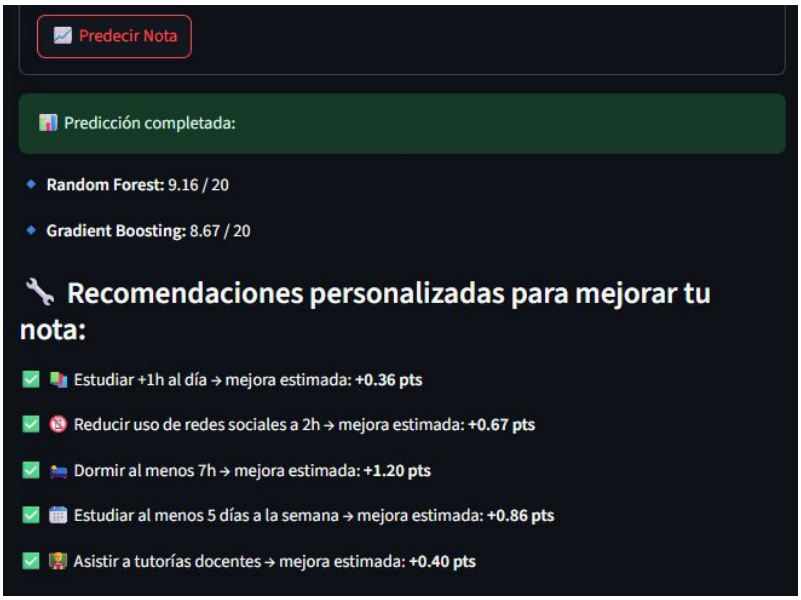

Figura 1.2: Predicción académica y recomendaciones generadas por la plataforma Streamlit

1.4 Objetivos de la Investigación

1.4.1 Objetivo General

- Desarrollar un modelo predictivo y un sistema de recomendaciones personalizadas basado en *aprendizaje automático* para mejorar la calificación académica de los estudiantes de Ingeniería de Software de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos (UNMSM).

1.4.2 Objetivos Específicos

- Recolectar y organizar un conjunto de datos reales de estudiantes que incluya variables psicoeducativas, socioeconómicas y conductuales.
- Entrenar modelos de *aprendizaje automático supervisado*, específicamente *Random Forest* y *Gradient Boosting*, para predecir la calificación semestral de los estudiantes.
- Desarrollar un sistema de recomendaciones personalizadas mediante simulaciones basadas en el impacto de modificar variables clave como horas de estudio, asistencia o sueño.
- Implementar una plataforma interactiva utilizando *Streamlit* que permita realizar predicciones y visualizar recomendaciones en tiempo real.
- Evaluar el desempeño de los modelos predictivos utilizando métricas de error y ajuste, tales como:
  - **MAE (Mean Absolute Error)**: Error absoluto medio.

- **RMSE (Root Mean Squared Error)**: Raíz del error cuadrático medio.
- **MedAE (Median Absolute Error)**: Error absoluto mediano.
- **SMAPE (Symmetric Mean Absolute Percentage Error)**: Error porcentual absoluto medio simétrico.
- R^2 **(Coeficiente de determinación)**: Grado de ajuste del modelo.

1.5 Importancia de la investigación

La presente investigación contribuye a la innovación en el ámbito educativo universitario mediante la aplicación de técnicas de *aprendizaje automático* como herramientas de apoyo a la toma de decisiones académicas. Al integrar modelos predictivos con sistemas de recomendaciones, se busca ofrecer a los estudiantes información personalizada y basada en evidencia que les permita mejorar sus calificaciones semestrales. Asimismo, proporciona a docentes y autoridades académicas una base para diseñar estrategias pedagógicas orientadas al acompañamiento académico individualizado.

1.6 Justificación

1.6.1 Justificación teórica

Diversas investigaciones han demostrado la eficacia de los modelos de *aprendizaje automático supervisado* en la predicción del rendimiento académico a partir de variables heterogéneas Nizar et al. (2024); Oyedeji et al. (2020); Yang et al. (2025); Ahmed (2024). Algoritmos como *Random Forest* y *Gradient Boosting* han superado en varios casos a

métodos estadísticos tradicionales en cuanto a precisión y capacidad de generalización Hashim et al. (2020); Wang and Yu (2025); Buenaño-Fernández et al. (2019).

Sin embargo, la mayoría de estos enfoques se limitan a realizar predicciones sin proveer retroalimentación útil para el estudiante. En ese sentido, la presente investigación se justifica teóricamente al integrar un sistema de recomendaciones personalizadas, lo que representa un aporte adicional al enfoque tradicional de predicción, alineándose con propuestas recientes que buscan empoderar al estudiante a través de estrategias accionables basadas en datos Atalla et al. (2023); Alsubaie (2023); Nachouki et al. (2023).

1.6.2 Justificación práctica

La implementación de un sistema predictivo y de recomendaciones personalizadas tiene un alto valor práctico al ofrecer a los estudiantes de Ingeniería de Software de la UNMSM una herramienta interactiva que les permita anticipar su calificación semestral y tomar decisiones informadas para mejorarla. Esto puede impactar positivamente en la reducción de tasas de desaprobación, deserción y retraso académico. Pan and Dai (2024); Zhu et al. (2025); Alsubhi et al. (2023); Nizar et al. (2024); Hashim et al. (2020)

1.7 Limitaciones

- Una de las principales limitaciones de esta investigación es la disponibilidad y calidad del conjunto de datos utilizado. Se trabajó con una muestra limitada de datos académicos recolectados de los estudiantes de Ingeniería de Software de la UNMSM, lo que podría afectar la capacidad de generalizar los resultados a otros contextos o facultades. Aunque se aplicaron técnicas de preprocesamiento y validación,

la falta de datos de mayor diversidad y volumen puede influir en la precisión de los modelos predictivos desarrollados.

- Otra limitación importante radica en la selección de algoritmos de aprendizaje automático. Si bien se emplearon modelos robustos y ampliamente validados como *Random Forest* y *Gradient Boosting*, no se exploraron enfoques más avanzados como redes neuronales profundas o modelos híbridos, los cuales podrían ofrecer mejores resultados pero requieren mayores recursos computacionales y bases de datos más extensas.

- Además, la personalización de recomendaciones fue diseñada con base en criterios derivados de la importancia de características predichas por los modelos, sin incorporar sistemas avanzados de recomendación colaborativa o basados en contenido. Esta simplificación permite una implementación inicial funcional, pero limita el grado de adaptabilidad y profundidad de las recomendaciones ofrecidas.

- Finalmente, el modelo no contempla variables psicológicas o motivacionales que también influyen en el rendimiento académico. La omisión de estos factores se debió a la dificultad de recolectar datos sensibles y al enfoque centrado en información académica y cuantificable.

A pesar de estas limitaciones, esta investigación proporciona una base inicial sólida para el desarrollo de herramientas inteligentes orientadas a mejorar el rendimiento académico mediante técnicas de *aprendizaje automático*. Los resultados obtenidos pueden ser valiosos para futuras investigaciones que deseen incorporar variables adicionales, explorar algoritmos más sofisticados o extender el sistema a otros programas educativos.

1.8 Metas

- Desarrollar un modelo predictivo de calificación académica con alto rendimiento en términos de precisión, utilizando algoritmos de *aprendizaje automático* supervisado.

- Implementar un sistema de recomendaciones personalizadas que permita a los estudiantes explorar, mediante simulaciones interactivas, cómo ciertos cambios en sus hábitos académicos y conductuales podrían mejorar su calificación estimada.

- Contribuir al desarrollo de herramientas tecnológicas inteligentes aplicadas a la educación superior, promoviendo una toma de decisiones basada en datos tanto por parte de los estudiantes como de las instituciones educativas.

- Proporcionar una base metodológica y tecnológica que sirva como referencia para futuras investigaciones sobre predicción del rendimiento académico y sistemas de apoyo al estudiante en contextos universitarios.

Capítulo II

Revisión de la Literatura

2.1 Marco Teórico

2.1.1 Aprendizaje Automático (Machine Learning)

El aprendizaje automático es una rama de la inteligencia artificial que permite a las computadoras aprender de los datos y tomar decisiones sin ser programadas explícitamente Sekeroglu et al. (2021). Se basa en el desarrollo de algoritmos capaces de identificar patrones y realizar predicciones a partir de datos históricos.

Entre las principales categorías del aprendizaje automático se encuentran:

- **Aprendizaje supervisado:** Utiliza conjuntos de datos etiquetados para entrenar modelos que puedan predecir resultados futuros. Los algoritmos buscan una función que mapee entradas a salidas conocidas Hashim et al. (2020).

- **Aprendizaje no supervisado:** Se basa en datos sin etiquetar. El objetivo es encontrar patrones ocultos o estructuras subyacentes en los datos, como agrupamientos o

asociaciones Zhao et al. (2023).

- **Aprendizaje por refuerzo:** Consiste en entrenar agentes para que aprendan mediante la interacción con un entorno, optimizando sus decisiones a través de recompensas o penalizaciones Sekeroglu et al. (2021).

Asimismo, el *Deep Learning*, una subcategoría del aprendizaje automático, ha ganado notoriedad por su capacidad para modelar relaciones complejas en grandes volúmenes de datos mediante redes neuronales profundas Jiao et al. (2022).

2.1.2 Predicción del Rendimiento Académico

La predicción del rendimiento académico se ha convertido en una de las principales aplicaciones del aprendizaje automático en el ámbito educativo. Su objetivo es anticipar el desempeño de los estudiantes a partir de variables académicas, demográficas, conductuales y socioeconómicas Oyedeji et al. (2020).

Modelos como Random Forest, Gradient Boosting Regressor, Support Vector Machines (SVM), redes neuronales y árboles de decisión han demostrado ser eficaces para predecir calificaciones, identificar estudiantes en riesgo de deserción o fracaso académico, y apoyar la toma de decisiones pedagógicas Wang and Yu (2025); Ahmed (2024); Balabied and Eid (2023).

Factores que afectan el rendimiento académico

Diversos estudios identifican múltiples factores que inciden en el rendimiento estudiantil:

- Historial académico (promedio ponderado, rendimiento en cursos previos).

- Asistencia a clases y participación en entornos virtuales.
- Perfil socioeconómico.
- Edad, género y otras variables demográficas.
- Autoevaluaciones, hábitos de estudio y nivel de motivación Zhu et al. (2025); Nizar et al. (2024).

Según Alsubhi et al. (2023), una adecuada selección y procesamiento de estas variables permite alcanzar modelos con altos niveles de precisión (mayores al 80%) en contextos universitarios.

2.1.3 Sistemas de Recomendación Personalizados

Los sistemas de recomendación son tecnologías diseñadas para sugerir ítems o acciones personalizadas a los usuarios, basándose en sus características, preferencias o comportamientos previos Albreiki et al. (2021). En el ámbito educativo, se utilizan para recomendar materiales de estudio, estrategias de aprendizaje, cursos complementarios, entre otros.

Los principales enfoques en sistemas de recomendación incluyen:

- **Filtrado colaborativo:** Se basa en las preferencias de usuarios similares para generar recomendaciones.
- **Basado en contenido:** Analiza las características de los ítems que el usuario ha consumido previamente para recomendar otros similares.
- **Modelos híbridos:** Combinan filtrado colaborativo y basado en contenido para mejorar la precisión de las recomendaciones Alsubaie (2023).

Según Atalla et al. (2023), la incorporación de sistemas de recomendación en entornos educativos puede mejorar la motivación del estudiante, facilitar el aprendizaje autónomo y, en consecuencia, elevar el rendimiento académico.

2.1.4 Aplicaciones del Aprendizaje Automático en Educación Superior

En los últimos años, las instituciones de educación superior han incorporado técnicas de aprendizaje automático para optimizar procesos administrativos y pedagógicos. En particular, en el contexto de carreras como Ingeniería de Software, se han desarrollado modelos para:

- Predecir el rendimiento académico y la deserción estudiantil Buenaño-Fernández et al. (2019).
- Identificar factores de riesgo y generar alertas tempranas.
- Evaluar competencias técnicas mediante análisis automatizado de código o tareas Nizar et al. (2024).
- Sugerir contenidos formativos de forma personalizada, mediante sistemas adaptativos de aprendizaje Atalla et al. (2023).

2.2 Marco Conceptual

2.2.1 Aprendizaje Automático

El *aprendizaje automático* (*Machine Learning*) es una rama de la inteligencia artificial que permite a los sistemas aprender y mejorar su rendimiento de manera automática a partir de la experiencia, sin ser programados explícitamente Ahmed (2024); Zhao et al. (2023). Los algoritmos supervisados, como *Random Forest* y *Gradient Boosting*, son ampliamente utilizados en la predicción de resultados en contextos educativos debido a su capacidad para manejar grandes volúmenes de datos y generar modelos explicativos y predictivos de alto rendimiento Pan and Dai (2024); Zhu et al. (2025); Alsubhi et al. (2023); Nachouki et al. (2023).

2.2.2 Random Forest

El algoritmo *Random Forest* es un modelo de ensamblado basado en árboles de decisión que construye múltiples árboles durante el entrenamiento y produce la media de sus predicciones para mejorar la precisión y reducir el sobreajuste Nizar et al. (2024); Balabied and Eid (2023); Chen and Liu (2024). Su robustez y facilidad de interpretación lo convierten en una herramienta eficaz para la predicción del rendimiento académico Pan and Dai (2024); Zhu et al. (2025); Nachouki et al. (2023).

2.2.3 Gradient Boosting

Gradient Boosting es una técnica de aprendizaje supervisado que construye modelos secuenciales, donde cada modelo intenta corregir los errores cometidos por los modelos

anteriores. Este método es especialmente eficaz cuando se trabaja con variables complejas y relaciones no lineales Jiao et al. (2022); Zhao et al. (2023); Selvakumari et al. (2023).

2.2.4 Predicción del Rendimiento Académico

La predicción del rendimiento académico busca estimar las calificaciones futuras de los estudiantes a partir de datos históricos, psicológicos, sociales y comportamentales. Este tipo de modelos tiene múltiples aplicaciones: detección temprana de riesgo académico, personalización de la enseñanza y orientación vocacional Hashim et al. (2020); Oyedeji et al. (2020); Namoun and Alshanqiti (2020); Albreiki et al. (2021). Estudios recientes validan el uso del aprendizaje automático como método eficaz para este fin Yang et al. (2025); Ahmed (2024); Guanin-Fajardo et al. (2024).

2.2.5 Sistemas de Recomendación en Educación

Los *sistemas de recomendación personalizados* en el contexto educativo permiten generar sugerencias adaptadas a las características de cada estudiante. Al incorporar simulaciones basadas en modelos predictivos, se pueden ofrecer recomendaciones específicas sobre qué variables modificar (como horas de estudio, sueño, uso de redes sociales o asistencia) para mejorar el rendimiento académico Atalla et al. (2023); Delahoz-Dominguez and Hijón-Neira (2024); Alsubaie (2023).

2.2.6 Variables Predictoras del Desempeño Académico

Diversas investigaciones identifican variables que influyen significativamente en el rendimiento académico: características demográficas (edad, sexo), académicas (asistencia, créditos aprobados), socioeconómicas (ingreso familiar, transporte), tecnológicas (acceso a computadoras y smartphones), y personales (salud, habilidades, motivación) Alsubhi et al. (2023); Oyedeji et al. (2020); Buenaño-Fernández et al. (2019); Sekeroglu et al. (2021).

2.2.7 Aplicaciones Interactivas con Streamlit

Streamlit es una biblioteca de Python que permite construir aplicaciones web interactivas para visualización y análisis de datos. En el presente estudio, se utiliza Streamlit para integrar el modelo predictivo y ofrecer predicciones en tiempo real junto con recomendaciones personalizadas, facilitando su uso por parte de estudiantes y docentes sin necesidad de conocimientos técnicos avanzados **?**.

2.2.8 Importancia del Apoyo a la Toma de Decisiones Académicas

El uso de herramientas basadas en inteligencia artificial permite transformar datos educativos en conocimiento accionable. Estas herramientas empoderan a los estudiantes para que tomen decisiones informadas sobre su proceso de aprendizaje, y permiten a las instituciones diseñar intervenciones más efectivas para mejorar el rendimiento académico general Jiao et al. (2022); Delahoz-Dominguez and Hijón-Neira (2024); Atalla et al. (2023).

2.3 Antecedentes

2.3.1 Predicción del rendimiento académico y su importancia

La predicción del rendimiento académico es una tarea clave en el campo de la analítica del aprendizaje, ya que permite identificar factores que influyen en el desempeño estudiantil y proponer estrategias de mejora. Su aplicación se ha expandido en diversos niveles educativos gracias a la disponibilidad de datos educativos y al desarrollo de técnicas avanzadas de aprendizaje automático. El uso de modelos predictivos permite anticipar el rendimiento futuro de un estudiante a partir de variables históricas, psicopedagógicas, socioeconómicas y comportamentales, lo cual brinda oportunidades para intervenir de forma temprana y personalizada Namoun and Alshanqiti (2020).

Diversas investigaciones han demostrado que el rendimiento académico está influenciado por múltiples factores, tales como las horas de estudio, el nivel de asistencia, el uso de plataformas educativas, la calidad del sueño, el entorno familiar y el nivel socioeconómico Zhao et al. (2023). Esto ha motivado el desarrollo de modelos predictivos que integran dichas variables para estimar calificaciones o identificar estudiantes en riesgo.

Modelos basados en aprendizaje supervisado, como Random Forest, Gradient Boosting, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN) y redes neuronales, han sido ampliamente utilizados para esta tarea. Entre ellos, los algoritmos basados en árboles, como Random Forest y Gradient Boosting, han destacado por su capacidad para manejar variables heterogéneas, interpretar la importancia de las características y ofrecer un buen rendimiento general Chen and Liu (2024).

Los modelos predictivos no solo permiten estimar calificaciones, sino también realizar segmentaciones del alumnado, identificar patrones de abandono, detectar brechas

de aprendizaje y proponer intervenciones pedagógicas personalizadas. En este sentido, se ha incrementado el interés en desarrollar plataformas que integren estos modelos en tiempo real, para apoyar la toma de decisiones académicas tanto de los estudiantes como de los docentes Wang and Yu (2025).

2.3.2 Sistemas de recomendación personalizados en educación

Los sistemas de recomendación personalizados han evolucionado significativamente en el ámbito educativo, con el objetivo de ofrecer sugerencias adaptadas al perfil de cada estudiante. Estos sistemas, que inicialmente se aplicaron en plataformas de comercio electrónico o entretenimiento, ahora se utilizan para proponer recursos de aprendizaje, estrategias de estudio y rutas de mejora académica Delahoz-Dominguez and Hijón-Neira (2024).

En el contexto educativo, los sistemas de recomendación pueden clasificarse en tres tipos principales: basados en contenido, colaborativos y basados en conocimiento. Los sistemas basados en contenido analizan las características del estudiante y los recursos disponibles para generar recomendaciones. Los colaborativos se basan en la experiencia y decisiones de otros estudiantes similares. Los basados en conocimiento integran reglas o modelos pedagógicos explícitos que guían las sugerencias Atalla et al. (2023).

Una tendencia emergente es la combinación de modelos predictivos y sistemas de recomendación. Esta integración permite no solo estimar el rendimiento futuro, sino también simular escenarios hipotéticos en los que se modifican variables clave —como las horas de estudio, el uso de redes sociales, o la asistencia a clase— para observar su impacto en la calificación proyectada. Esto ofrece a los estudiantes una herramienta proactiva para mejorar su desempeño mediante la toma de decisiones informadas Ahmed (2024).

En investigaciones recientes se ha evidenciado el uso de plataformas interactivas, como aquellas desarrolladas con herramientas como Streamlit o Dash, que permiten visualizar predicciones y recomendaciones personalizadas en tiempo real. Estas plataformas aumentan el compromiso del estudiante y brindan una experiencia más personalizada y centrada en el usuario Zhu et al. (2025).

2.3.3 Aplicación de aprendizaje automático en el análisis educativo

El aprendizaje automático ha permitido construir modelos robustos para la predicción de resultados académicos, clasificación de estudiantes y análisis de factores de éxito. En especial, los algoritmos de tipo *ensemble*, como Random Forest y Gradient Boosting, se han posicionado como una alternativa eficaz debido a su capacidad para reducir el sobreajuste y mejorar la precisión Pan and Dai (2024).

Además, el uso de técnicas de interpretabilidad de modelos, como SHAP (*SHapley Additive exPlanations*) o LIME (*Local Interpretable Model-agnostic Explanations*), ha permitido comprender mejor las decisiones de los modelos, facilitando su adopción por parte de instituciones educativas al brindar explicaciones claras sobre qué variables tienen mayor impacto en el rendimiento Nizar et al. (2024).

Por último, cabe resaltar que, si bien existen numerosos modelos predictivos desarrollados para contextos generales, pocos se han adaptado específicamente a las características del estudiantado de universidades públicas latinoamericanas, como la UNMSM. Esta tesis busca contribuir a este vacío, proponiendo un modelo específico para estudiantes de Ingeniería de Software, con datos reales del contexto local, y complementándolo con un sistema de recomendaciones que simula el impacto de decisiones concretas en la mejora del rendimiento académico.

2.4 Estado del Arte

El uso de técnicas de aprendizaje automático para predecir el rendimiento académico ha sido un campo de investigación en constante crecimiento, dado su potencial para apoyar a estudiantes e instituciones en la toma de decisiones informadas. Existen diversas aproximaciones metodológicas para abordar esta problemática, que van desde modelos estadísticos tradicionales hasta algoritmos avanzados de aprendizaje supervisado y sistemas de recomendación.

Los enfoques tradicionales para predecir el rendimiento académico se han basado en regresiones lineales, análisis discriminante y árboles de decisión simples Namoun and Alshanqiti (2020). Sin embargo, estos modelos presentan limitaciones para capturar relaciones no lineales complejas entre las variables psicoeducativas, socioeconómicas y conductuales que influyen en el rendimiento de los estudiantes. En este contexto, el aprendizaje automático ha emergido como una alternativa efectiva, al permitir modelar patrones complejos en grandes volúmenes de datos.

Entre los algoritmos más utilizados se encuentran los modelos basados en árboles, como el Random Forest y el Gradient Boosting, los cuales han demostrado un rendimiento superior en tareas de regresión y clasificación en educación Albreiki et al. (2021). Estos modelos destacan por su capacidad de manejar datos heterogéneos, tolerar valores atípicos y ofrecer interpretabilidad parcial mediante la importancia de variables. Investigaciones previas han utilizado Random Forest para predecir el promedio ponderado de estudiantes universitarios con resultados prometedores, alcanzando coeficientes de determinación (R^2) superiores a 0.75 Pan and Dai (2024); Zhu et al. (2025); Alsubhi et al. (2023); Hashim et al. (2020); Nachouki et al. (2023). Por su parte, el Gradient Boosting, debido a su enfoque secuencial de optimización, ha mostrado una capacidad superior

para minimizar el error cuadrático medio, especialmente en datasets con ruido o variables redundantes Ahmed (2024); Zhao et al. (2023).

En cuanto a las variables utilizadas como entrada en los modelos predictivos, se ha evidenciado que los factores académicos tradicionales (como el historial de calificaciones) no son suficientes. Diversas investigaciones han incorporado variables conductuales y psicoeducativas, tales como las horas de estudio, calidad del sueño, hábitos digitales, motivación intrínseca y asistencia a clases Nizar et al. (2024); Oyedeji et al. (2020); Wang and Yu (2025); Buenaño-Fernández et al. (2019), así como también indicadores socioeconómicos, como el nivel de ingreso familiar o el acceso a recursos tecnológicos Balabied and Eid (2023); Delahoz-Dominguez and Hijón-Neira (2024); Guanin-Fajardo et al. (2024). Estas variables permiten una comprensión más holística del proceso de aprendizaje, lo cual resulta fundamental para diseñar intervenciones personalizadas.

En cuanto al desarrollo de sistemas interactivos que integren modelos predictivos y generen recomendaciones personalizadas, aún son limitadas las investigaciones que aborden ambos componentes de manera conjunta. Algunos trabajos han propuesto dashboards que visualizan riesgos de deserción Sekeroglu et al. (2021) o plataformas de tutoría inteligente que sugieren recursos de estudio Atalla et al. (2023). Sin embargo, la combinación de un modelo de predicción con un sistema que simule escenarios hipotéticos —por ejemplo, aumentando las horas de estudio o reduciendo el tiempo en redes sociales— para observar su impacto en la nota estimada, representa una innovación significativa, alineada con el paradigma de educación personalizada impulsada por la inteligencia artificial Yang et al. (2025); Jiao et al. (2022); Alsubaie (2023).

Finalmente, el uso de plataformas ligeras y accesibles como Streamlit ha facilitado la integración de modelos predictivos en interfaces interactivas que no requieren conocimien-

tos técnicos por parte de los usuarios. Esto representa un avance importante para la democratización del análisis educativo basado en datos Selvakumari et al. (2023).

En síntesis, la literatura revisada respalda la relevancia de utilizar modelos avanzados de aprendizaje automático para la predicción del rendimiento académico, y destaca la necesidad de sistemas que no solo diagnostiquen, sino también orienten proactivamente a los estudiantes mediante recomendaciones personalizadas basadas en simulaciones de variables clave. Esta investigación busca contribuir en ambos frentes: mediante el desarrollo de un modelo predictivo robusto y un sistema interactivo de apoyo a la toma de decisiones académicas.

Capítulo III

Hipótesis

Este capítulo presenta la hipótesis general y las hipótesis específicas que orientan la validación del modelo predictivo y sistema de recomendaciones personalizadas desarrollado para mejorar el rendimiento académico de los estudiantes de Ingeniería de Software de la UNMSM.

3.1 Hipótesis General

El uso de modelos de aprendizaje automático supervisado, específicamente Random Forest y Gradient Boosting, combinado con un sistema de recomendaciones personalizadas, permite predecir con precisión la calificación académica semestral y proponer acciones concretas que mejoren el rendimiento académico de los estudiantes de Ingeniería de Software de la UNMSM.

3.2 Hipótesis Específicas

- **Hipótesis 1:** Las variables psicoeducativas, socioeconómicas y conductuales (como horas de estudio, sueño, uso de redes sociales y asistencia) son predictoras significativas del rendimiento académico de los estudiantes. *Esta hipótesis se fundamenta en investigaciones previas que demuestran la influencia significativa de factores conductuales y contextuales sobre el desempeño académico, como lo señalan Pan y Dai (2024), y Alsubhi et al. (2023), quienes emplearon algoritmos de aprendizaje supervisado para evaluar la importancia de características como el entorno familiar, hábitos de estudio y asistencia.* Pan and Dai (2024) Alsubhi et al. (2023)

- **Hipótesis 2:** El modelo de *Gradient Boosting Regressor* presenta un mejor desempeño que el modelo *Random Forest Regressor* en la predicción de calificaciones académicas, según las métricas de MAE, RMSE, MedAE, SMAPE y R^2.
  Estudios recientes han comparado algoritmos de aprendizaje automático en tareas de predicción académica, mostrando que el modelo Gradient Boosting puede superar en precisión al modelo Random Forest, especialmente al optimizar hiperparámetros y realizar una adecuada selección de características. Este mejor desempeño se observa en métricas como el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE), el Error Absoluto Mediano (MedAE), el Error Porcentual Absoluto Simétrico (SMAPE) y el coeficiente de determinación (R^2). Jiao et al. (2022) Zhao et al. (2023) Chen and Liu (2024)

- **Hipótesis 3:** El sistema de recomendaciones personalizadas, basado en simulaciones de escenarios modificando variables clave, permite identificar estrategias viables para mejorar el rendimiento académico individual de los estudiantes.

La literatura respalda la efectividad de los sistemas inteligentes de recomendación académica que personalizan estrategias según el perfil del estudiante. Por ejemplo, Atalla et al. (2023) y Delahoz-Dominguez y Hijón-Neira (2024) muestran cómo estos sistemas pueden orientar eficazmente decisiones académicas mediante simulación y análisis predictivo Delahoz-Dominguez and Hijón-Neira (2024) Atalla et al. (2023).

- **Hipótesis 4:** La integración del modelo predictivo en una plataforma interactiva facilita la interpretación de resultados y la toma de decisiones informadas por parte de los estudiantes y orientadores académicos.
  La investigación de Wang y Yu (2025) y Oyedeji et al. (2020) muestra que la visualización interactiva de resultados predictivos mejora la comprensión y promueve el uso de herramientas de IA por parte de usuarios no técnicos en contextos educativos, facilitando decisiones informadas Oyedeji et al. (2020) Wang and Yu (2025).

3.3 Variables

3.3.1 Variables Independientes

Las variables independientes son aquellas características que se utilizaron como entrada para los modelos predictivos. Estas se clasifican en tres grupos principales:

- **Variables psicoeducativas:** Relacionadas con hábitos y comportamientos que influyen directamente en el rendimiento académico. Incluyen:
  - **Horas de estudio por día:** de 0 a 13 horas.
  - **Frecuencia de estudio semanal:** de 0 a 7 días.

  - **Modalidad de aprendizaje:** presencial (*offline*) u online.
  - **Horas de sueño diarias:** de 5 a 10 horas.
  - **Nivel de asistencia a clases:** porcentaje entre 0 y 100%.
  - **Nivel de inglés:** básico, intermedio o avanzado.
  - **Participación en tutorías docentes:** sí o no.
  - **Habilidades técnicas declaradas:** programación, desarrollo web, IA, ciberseguridad, etc.
  - **Horas dedicadas a desarrollar habilidades:** de 0 a 12 horas por día.
  - **Riesgo académico:** sí o no.
  - **Antecedentes de suspensión académica:** sí o no.

- **Variables conductuales:** Reflejan el estilo de vida y la gestión del tiempo del estudiante. Se consideran:
  - **Uso de redes sociales:** de 0 a 20 horas por día.
  - **Participación en actividades extracurriculares:** sí o no.
  - **Horas de ejercicio semanal:** de 2 a 6 horas.
  - **Transporte universitario:** tipo (público, propio o mixto) y **tiempo de traslado** (3 a 7 horas semanales).
  - **Horas de trabajo semanal:** de 0 a 35 horas.

- **Variables socioeconómicas y contextuales:** Representan el entorno familiar y los recursos disponibles del estudiante. Incluyen:
  - **Año de ingreso a la universidad:** entre 2013 y 2023.

- **Año de egreso de secundaria:** entre 2012 y 2022.
- **Edad:** entre 18 y 27 años.
- **Sexo:** masculino o femenino.
- **Semestre actual:** del 1 al 24.
- **Recepción de beca por rendimiento:** sí o no.
- **Ingreso familiar anual:** de 4,000 a 2,000,000 soles.
- **Créditos académicos completados:** de 0 a 120.
- **Acceso a dispositivos tecnológicos:** smartphone y computadora personal (sí o no).
- **Convivencia:** vive con familiares o solo (categorías: *family, bachelor*).
- **Estado sentimental:** soltero, en una relación (formal e informal), casado y comprometido.
- **Problemas de salud o discapacidad física:** sí o no.
- **Horas de clases semanales:** de 0 a 36 horas.

3.3.2 Variable Dependiente

- **Calificación académica semestral final:** Variable objetivo del estudio. Corresponde a la nota final obtenida por el estudiante al término del semestre académico, expresada en una escala de 0 a 20. Esta variable es continua y representa el rendimiento académico general.

Capítulo IV

Materiales y Métodos

4.1 Diseño de la Investigación

El presente estudio empleó un diseño de investigación de tipo *cuasi-experimental aplicado con enfoque cuantitativo*, orientado a evaluar el desempeño de modelos de aprendizaje automático en la predicción del rendimiento académico de estudiantes de Ingeniería de Software de la UNMSM, así como la efectividad de un sistema de recomendaciones personalizadas. El objetivo principal fue determinar la precisión de los algoritmos Random Forest y Gradient Boosting al predecir las calificaciones finales, así como analizar cómo las recomendaciones generadas podrían contribuir a mejorar el desempeño académico mediante simulaciones interactivas.

El estudio se basó en un conjunto de datos reales recolectados de estudiantes universitarios, a través de un formulario en línea aplicado mediante *Google Forms*, incluyendo variables psicoeducativas, conductuales y socioeconómicas, las cuales sirvieron como insumo para entrenar, validar y evaluar los modelos predictivos.

4.2 Descripción de las Tareas

4.2.1 Tarea 1: Predicción de la Calificación Académica Final

La primera tarea consistió en entrenar y evaluar modelos de aprendizaje automático supervisado para predecir la calificación final del estudiante en una escala del 0 al 20. Esta predicción se realizó en función de múltiples variables independientes, tales como horas de estudio, calidad del sueño, nivel de asistencia, uso de redes sociales, nivel de ingresos, entre otras. El desempeño de los modelos fue evaluado utilizando métricas estándar como el Coeficiente de Determinación (R^2), el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE), el Error Absoluto Mediano (MedAE) y el Porcentaje de Error Absoluto Medio Simétrico (SMAPE).

4.2.2 Tarea 2: Generación de Recomendaciones Personalizadas

La segunda tarea implicó el desarrollo de un sistema de simulación que, a partir del modelo predictivo, pudiera generar recomendaciones personalizadas. Este sistema permite modificar de manera controlada los valores de variables clave (como horas de estudio o sueño) para observar el impacto proyectado en la nota estimada del estudiante. Estas simulaciones se presentan mediante una interfaz interactiva implementada en Streamlit, con el fin de facilitar la comprensión de las recomendaciones y fomentar la toma de decisiones informadas.

4.3 Herramientas de Desarrollo

Para el desarrollo de la solución se utilizó el lenguaje de programación Python, junto con librerías especializadas en ciencia de datos y machine learning como:

- **Pandas** y **NumPy** para la manipulación y análisis de datos.
- **Scikit-learn** para la implementación de los algoritmos *Random Forest* y *Gradient Boosting*, así como para la división de datos y evaluación del rendimiento del modelo.
- **Joblib** para la serialización y carga de los modelos entrenados.
- **Streamlit** para el desarrollo de una plataforma interactiva que permite ingresar datos, ejecutar los modelos predictivos y mostrar recomendaciones personalizadas para mejorar la calificación académica.

Todo el proceso de desarrollo, desde la preparación del conjunto de datos hasta la implementación del sistema interactivo, se llevó a cabo en Google Colab, sin necesidad de aceleración por GPU, dado que los modelos empleados no requerían un alto poder computacional. Para el despliegue de la aplicación, se utilizó Streamlit, la cual fue conectada a un repositorio de GitHub que contiene todo el código fuente, permitiendo su ejecución en un entorno abierto y accesible.

4.4 Recolección de Datos

Para la ejecución del presente estudio, se recopiló un conjunto de datos compuesto por **1,194 registros** correspondientes a estudiantes de la Escuela Profesional de Ingeniería

de Software de la Universidad Nacional Mayor de San Marcos (UNMSM). El objetivo principal fue entrenar modelos predictivos y desarrollar un sistema de recomendaciones personalizadas que contribuyan a mejorar el rendimiento académico.

Los datos fueron obtenidos mediante encuestas aplicadas de forma voluntaria, complementadas con registros académicos institucionales anonimizados, en cumplimiento con las directrices éticas de confidencialidad y uso responsable de la información. La recolección incluyó estudiantes de diversos semestres y años de ingreso, lo cual proporcionó una base heterogénea y representativa del perfil estudiantil.

4.4.1 Tarea 1: Predicción de la Calificación Académica Semestral Final

En esta primera tarea se formuló un problema de regresión supervisada, cuya variable objetivo fue la **nota promedio del semestre** (en escala 0 a 20). Para ello, se consideraron variables académicas, personales, socioeconómicas y contextuales que, según estudios previos, tienen impacto significativo en el desempeño estudiantil. Entre las variables más relevantes se encuentran:

- Año de ingreso, edad y semestre actual
- Asistencia promedio a clases (%)
- Horas de estudio diarias y frecuencia de estudio semanal
- Uso de redes sociales (horas por día)
- Nivel de inglés

- Ingreso familiar anual
- Acceso a computadora personal y smartphone
- Participación en tutorías y actividades extracurriculares
- Horas de sueño, transporte y trabajo semanal

Se realizó un proceso de limpieza, normalización y codificación de los datos para su posterior uso en algoritmos de aprendizaje automático. Además, se evaluaron correlaciones y distribuciones para detectar valores atípicos o inconsistencias.

4.4.2 Tarea 2: Generación de Recomendaciones Personalizadas

A partir del mismo conjunto de datos, se diseñó un sistema de simulación inteligente que permite al estudiante modificar ciertas variables de entrada (como horas de estudio, sueño o tiempo en redes sociales) y visualizar el impacto proyectado en su nota final estimada. Este sistema interactivo fue implementado con **Streamlit** y se apoya directamente en los modelos predictivos entrenados (*Random Forest* y *Gradient Boosting*) para generar recomendaciones personalizadas.

En lugar de aplicar técnicas de *data augmentation*, se desarrolló una estrategia basada en análisis contrafactuales o simulaciones tipo *what-if*, donde se ajustan selectivamente variables académicas, conductuales y de estilo de vida del perfil actual del estudiante. Cada cambio se evalúa en tiempo real mediante el modelo entrenado, lo que permite identificar intervenciones específicas que podrían mejorar significativamente su rendimiento estimado.

Tabla 4.1: Características del Conjunto de Datos por Tarea

Características	**Tarea 1: Predicción**	**Tarea 2: Recomendación**
Número total de registros	1,194	1,194
Tipo de salida	Regresión (nota 0–20)	Estimación simulada (*what-if*)
Número de variables	34	Subconjunto editable (8–10)
Fuentes de datos	Encuestas + registros académicos	Mismo dataset + simulaciones contrafactuales
Técnicas de aumento	No aplicadas	No (se simulan posibles cambios en variables)
Herramienta de despliegue	No aplica	Plataforma interactiva en **Streamlit**

4.5 Preprocesamiento de los Datos

4.5.1 Dataset

El conjunto de datos utilizado en esta investigación está conformado por **1,194 registros** de estudiantes de la Escuela Profesional de Ingeniería de Software de la Universidad Nacional Mayor de San Marcos (UNMSM). Cada registro representa información correspondiente a un estudiante en un semestre específico, incluyendo variables académicas, personales, contextuales y socioeconómicas. El objetivo del preprocesamiento fue garantizar la calidad, consistencia y adecuación de los datos para su posterior análisis mediante algoritmos de aprendizaje automático, específicamente **Random Forest** y **Gradient Boosting**.

Figura 4.1: Proceso de Preprocesamiento de Datos Académicos para Modelos Predictivos

4.5.2 Limpieza y Tratamiento de Inconsistencias

Se realizó una depuración inicial del conjunto de datos para eliminar:

- Registros duplicados.
- Errores ortográficos frecuentes (e.g., “dvance” por “advanced”, “Syber” por “Cyber”).
- Categorías incoherentes o sin valor analítico (e.g., “No skill”, “Nothing properly”).
- Rótulos redundantes o mal estandarizados (e.g., “Graphics design”, “Graphics Designing”).

4.5.3 Imputación de Valores Faltantes

Se identificaron valores nulos o faltantes en múltiples variables del conjunto de datos. Para evitar perder registros valiosos y garantizar la consistencia del dataset, se aplicaron técnicas de imputación diferenciadas según el tipo de variable:

- **Variables categóricas**: se reemplazaron los valores faltantes utilizando la *moda* (valor más frecuente). Esto se aplicó a variables como *estado_sentimental*, *modalidad_aprendizaje*, y *nivel_ingles*. Por ejemplo, más del 50% de los estudiantes

reportaban estar “Soltero” y un registro carecía de ese dato, se imputaba como “Soltero”.

- **Variables numéricas**: se utilizó la *mediana* como estrategia de imputación, ya que este estadístico es robusto ante valores atípicos que podrían distorsionar la media. Esta técnica se aplicó a variables como *horas_redes_sociales*, *horas_estudio_dia*, e *ingreso_familiar_anual*. Por ejemplo, si la mediana de *horas_estudio_dia* era 3.5 y había registros faltantes, estos se completaban con ese valor.

Esta imputación fue crítica para mantener la integridad del dataset, especialmente considerando que el tamaño total de la muestra era de 1,194 registros, y descartar filas podría haber afectado la representatividad de los modelos entrenados.

4.5.4 Transformación y Estandarización de Variables

Para facilitar el análisis automático, se realizó la transformación de varias variables:

- **Binarias textuales** (e.g., *sí/no*, *online/offline*) se codificaron como 1 y 0.
- **Rangos textuales** (e.g., “0 a 13”) fueron reemplazados por su media aproximada (e.g., 6.5).
- **Campos múltiples**: variables como *habilidades* y *área_interes*, que presentaban respuestas abiertas y diversas, fueron transformadas asignando un identificador numérico único a cada categoría distinta. Para ello:
  - Se ordenaron alfabéticamente todas las respuestas distintas encontradas.
  - A cada categoría se le asignó un valor entero consecutivo, desde 1 hasta n, donde n es el número total de categorías.

Esta estrategia permitió mantener la unicidad de cada entrada sin considerar similitud semántica entre ellas.

4.5.5 Codificación de Variables Categóricas

Con el fin de hacer las variables categóricas interpretables por los modelos de aprendizaje automático, se aplicaron técnicas de transformación numérica a las siguientes variables:

- **Variables nominales**: estas variables no tienen un orden inherente entre sus categorías (por ejemplo, *nivel_ingles*, *modalidad_aprendizaje*, *estado_sentimental*). Se utilizó *Label Encoding*, que asigna un número entero distinto a cada categoría. Por ejemplo:
  - *nivel_ingles*: `básico` $\rightarrow$ 0, `intermedio` $\rightarrow$ 1, `avanzado` $\rightarrow$ 2.
  - *modalidad_aprendizaje*: `online` $\rightarrow$ 0, `offline` $\rightarrow$ 1.
  - *estado_sentimental*: `Comprometido` $\rightarrow$ 1, `En una relación` $\rightarrow$ 2, `Casado` $\rightarrow$ 3, `Relación informal` $\rightarrow$ 4, y `Soltero` $\rightarrow$ 5.

  Esta codificación fue elegida por su simplicidad y porque los algoritmos utilizados (Random Forest y Gradient Boosting) no se ven afectados negativamente por la escala de los valores.

4.5.6 Detección y Corrección de Valores Atípicos

Se revisaron las variables numéricas para identificar outliers. Se aplicaron reglas de umbral basadas en valores posibles y coherentes:

- *horas_estudio_dia*: limitado entre 0 y 13.

- *horas_sueno_dia*: restringido entre 5 y 10.
- *ingreso_familiar_anual*: normalizado entre 4,000 y 2,000,000 soles.

Registros con valores fuera del rango lógico fueron ajustados o eliminados según el caso.

4.5.7 Validación de Rango y Consistencia

Se realizaron validaciones finales para asegurar que todas las variables cumplieran sus rangos esperados:

- *edad*: 18 a 27 años.
- *semestre_actual*: entre 1 y 24.
- *nota_semestre*: entre 0 y 20 (variable objetivo).
- *creditos_completados*, *horas_trabajo_semanal*, *horas_habilidades_dia*, entre otros.

4.5.8 Construcción del Dataset Final

Como resultado, se generó un conjunto de datos completamente estructurado, sin valores nulos, inconsistencias ni categorías erróneas. Este dataset se utilizó para dos tareas:

- **Tarea 1: Predicción de la Nota Final del Semestre.**
- **Tarea 2: Recomendaciones Personalizadas via Simulación Contrafactual.**

El dataset fue exportado en formato `.csv`, con codificación UTF-8, listo para su uso en modelos de aprendizaje automático.

$$\text{Shape final del dataset:} \quad X \in R^{1194\times 34}, \quad y \in R^{1194} \tag{4.1}$$

4.6 Arquitectura de los Modelos

En base al análisis del estado del arte, se seleccionaron algoritmos de aprendizaje automático reconocidos por su eficacia en tareas de predicción sobre datos tabulares: **Random Forest** y **Gradient Boosting**. Ambos modelos fueron entrenados utilizando el dataset académico procesado de estudiantes de Ingeniería de Software de la UNMSM. La elección de estos modelos responde a su capacidad para manejar variables categóricas y numéricas, su robustez ante valores atípicos, y su buen desempeño general en tareas de regresión y clasificación educativa.

El presente capítulo describe la arquitectura lógica del sistema predictivo y el sistema de recomendaciones personalizadas desarrollados. La arquitectura general se divide en dos módulos principales: el **Módulo de Preprocesamiento**, que transforma los datos brutos del estudiante en entradas válidas para el modelo, y el **Módulo de Aprendizaje Automático**, donde se encuentran los modelos entrenados y el motor de recomendaciones.

4.6.1 Módulo de Preprocesamiento

Este módulo recibe como entrada la información de un estudiante (ya sea desde el dataset o desde el formulario interactivo en la plataforma Streamlit). El preprocesamiento incluye:

- Imputación de valores faltantes.
- Codificación de variables categóricas.
- Escalamiento de variables numéricas (cuando es necesario).

- Alineación del formato de entrada con el usado durante el entrenamiento de los modelos.

Como resultado, se genera un vector de características de dimensiones $(1, 34)$ que representa el perfil académico, personal y contextual del estudiante. Este vector se utiliza como entrada directa para los modelos de predicción.

4.6.2 Módulo de Aprendizaje Automático

Este módulo contiene los modelos entrenados que permiten estimar el rendimiento académico del estudiante, así como generar recomendaciones personalizadas. Está dividido en dos partes: el **Modelo Predictivo** y el **Motor de Recomendaciones**.

Modelo Predictivo

Los modelos empleados son:

- **Random Forest Regressor:** Un conjunto de árboles de decisión entrenados sobre subconjuntos del dataset, cuya predicción final es el promedio de las predicciones de todos los árboles. Destaca por su capacidad de manejo de variables mixtas y su bajo riesgo de sobreajuste cuando se configura correctamente.
- **Gradient Boosting Regressor:** Construido de manera secuencial, donde cada árbol intenta corregir los errores del árbol anterior. Se seleccionó este modelo por su alta precisión en tareas de regresión con datos estructurados, y por su eficacia al capturar relaciones no lineales.

Ambos modelos fueron entrenados para predecir la **nota final del semestre** en una escala numérica continua. Se utilizaron métricas como el **MAE** (Mean Absolute Error),

RMSE (Root Mean Square Error), **MedAE** (Median Absolute Error), **SMAPE** (Symmetric Mean Absolute Percentage Error) y el $\mathbf{R^2}$ (Coeficiente de Determinación) para evaluar el rendimiento.

Motor de Recomendaciones Personalizadas

A partir del análisis de la importancia de las variables (feature importance) proporcionado por los modelos, se diseñó un sistema de recomendaciones que sugiere cambios específicos en los hábitos o condiciones del estudiante para mejorar su rendimiento académico proyectado.

Por ejemplo, si el modelo detecta que el tiempo en redes sociales y las horas de estudio son las variables más influyentes en su bajo rendimiento, el sistema sugerirá reducir el tiempo en redes sociales o aumentar las horas de estudio. Estas recomendaciones se generan dinámicamente mediante simulaciones controladas (what-if analysis) sobre el modelo entrenado.

En la figura 4.3 se muestra un diagrama general de la arquitectura implementada.

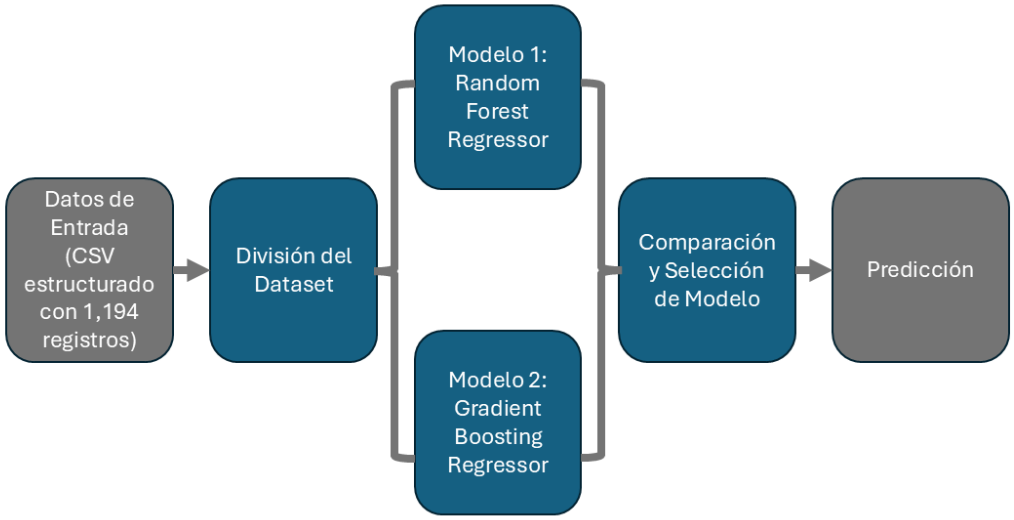

Figura 4.2: Arquitectura general del sistema predictivo

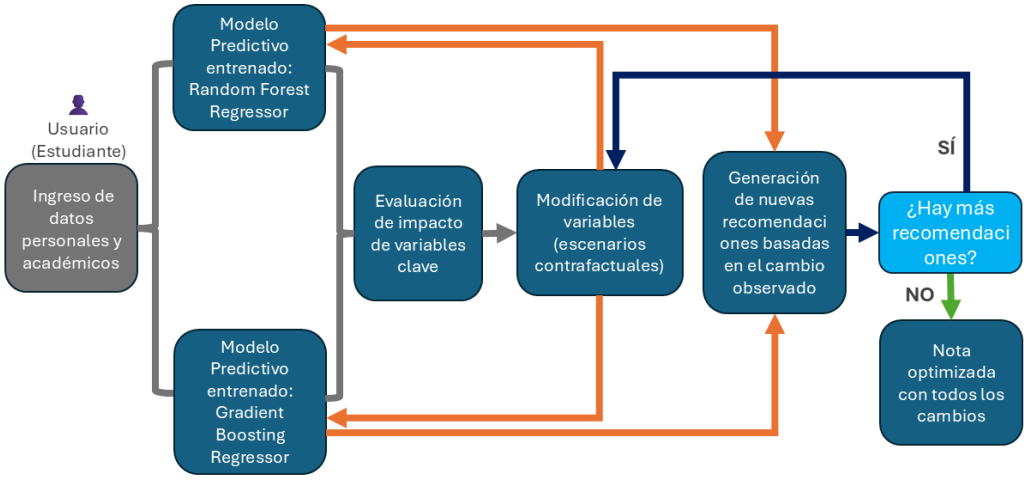

Figura 4.3: Arquitectura general del sistema recomendador basado en aprendizaje automático

Capa de Interpretabilidad

Ambos modelos (*Random Forest* y *Gradient Boosting*) fueron complementados con técnicas de interpretabilidad que permiten explicar el comportamiento interno de las predicciones, facilitando así una mayor confianza y utilidad práctica del sistema para usuarios no expertos.

- **Feature Importance (Importancia de Variables):** Este método muestra qué variables influyen más, en promedio, sobre la predicción del modelo. En este trabajo, se utilizó para identificar los factores globales más determinantes del rendimiento académico. Por ejemplo, las variables *horas de estudio*, *nivel de asistencia* y *tiempo en redes sociales* fueron las más relevantes en la predicción de la nota final según los modelos entrenados.

- **SHAP Values (Valores de SHAP):** SHAP (*SHapley Additive exPlanations*) permite descomponer cada predicción individual en las contribuciones positivas o negativas de cada variable. Esta técnica se basa en teoría de juegos y ofrece explicaciones consistentes para cada resultado. Por ejemplo, para un estudiante con una nota predicha de 14.2, SHAP puede mostrar que sus *horas de estudio elevadas* contribuyeron con +1.3 puntos a su predicción, mientras que su *alto tiempo en redes sociales* redujo su nota esperada en -0.7 puntos.

SHAP aplicado a un caso real:

- **Nota base del modelo:** 12.0
- **+1.2** por *alta asistencia*
- **+0.8** por *nivel de inglés avanzado*

- **-0.6** por *uso excesivo de redes sociales*
- **-0.2** por *pocas horas de sueño*
- **Resultado final: 13.2**

Estas herramientas permiten que el sistema no solo realice predicciones numéricas, sino que también ofrezca explicaciones comprensibles y accionables para los usuarios (estudiantes y docentes). Esto transforma el sistema en una plataforma transparente de apoyo a la toma de decisiones académicas, fundamentada en datos reales y fácilmente interpretables.

4.7 Entrenamiento

Para el entrenamiento de los modelos de aprendizaje automático, el conjunto de datos estudiantil fue dividido en una proporción de 80-20, utilizando el 80% de los registros para el entrenamiento y el 20% restante para la evaluación. Esta división permitió validar el rendimiento de los modelos sobre datos no vistos, asegurando su capacidad de generalización hacia nuevos estudiantes.

Se entrenaron dos modelos de regresión supervisada: **Random Forest Regressor** y **Gradient Boosting Regressor**, ambos pertenecientes a la familia de algoritmos basados en árboles de decisión. Estos modelos fueron seleccionados por su capacidad de manejar datos tabulares, capturar relaciones no lineales y ofrecer interpretabilidad a través de análisis de importancia de variables.

Durante el proceso de entrenamiento, los modelos fueron alimentados con datos estructurados que incluían variables académicas (por ejemplo, número de créditos, cursos

matriculados), variables personales (edad, uso de redes sociales, hábitos de estudio) y contextuales (ingresos familiares, transporte utilizado, participación en tutorías, entre otras). La variable objetivo fue la **calificación promedio del semestre**.

4.8 Evaluación

La evaluación de los modelos se realizó utilizando un enfoque cuantitativo y cualitativo, con el fin de medir la capacidad predictiva del sistema y su utilidad en contextos reales dentro de la Facultad de Ingeniería de Sistemas e Informática.

Además, se evaluó el sistema de **recomendaciones personalizadas** utilizando un enfoque cualitativo basado en escenarios reales. Se generaron recomendaciones específicas para estudiantes hipotéticos con diferentes perfiles (por ejemplo, estudiantes con baja asistencia, alto uso de redes sociales o bajo tiempo de estudio), evaluando si las sugerencias generadas eran coherentes y potencialmente útiles.

Esta evaluación integral permitió determinar la efectividad del modelo predictivo y del sistema de recomendación no solo desde una perspectiva técnica, sino también en términos de su aplicabilidad práctica para mejorar el rendimiento académico.

Los resultados cuantitativos de esta evaluación, así como los análisis de casos concretos, se presentan en el siguiente capítulo.

4.8.1 Métricas de Evaluación

Se presentan las métricas cuantitativas empleadas para evaluar el rendimiento de los modelos de regresión desarrollados. Estas métricas permiten medir la precisión de las predicciones realizadas respecto a las calificaciones reales de los estudiantes, ofreciendo una

visión integral del desempeño del sistema.

La selección de métricas se basó en literatura especializada en modelos de predicción educativa y aprendizaje automático, priorizando indicadores robustos ante valores atípicos y sensibles a errores grandes.

Las principales métricas utilizadas fueron:

- **MAE (Mean Absolute Error)**: Error absoluto promedio entre las predicciones y los valores reales. Es una medida sencilla y fácil de interpretar.
- **RMSE (Root Mean Squared Error)**: Raíz cuadrada del error cuadrático medio. Penaliza más fuertemente los errores grandes, siendo útil para detectar predicciones muy desviadas.
- **MedAE (Median Absolute Error)**: Mediana del error absoluto. Proporciona una medida robusta frente a valores atípicos, destacando el error típico en un caso promedio.
- **SMAPE (Symmetric Mean Absolute Percentage Error)**: Porcentaje de error absoluto simétrico. Mide el error relativo en términos porcentuales, útil para comparar el rendimiento en diferentes escalas.
- **R^2 (Coeficiente de determinación)**: Indica qué proporción de la varianza de la variable objetivo es explicada por el modelo. Un valor cercano a 1 representa un ajuste casi perfecto.

Estas métricas fueron calculadas para cada modelo entrenado (Random Forest y Gradient Boosting), permitiendo comparar objetivamente su capacidad de predicción. Los

resultados obtenidos orientan la elección del modelo más adecuado y fundamentan la generación de recomendaciones personalizadas con base en las predicciones realizadas.

4.8.2 Fase de Evaluación

La evaluación de los modelos se realizó utilizando el 20% del conjunto de datos, previamente reservado como conjunto de prueba. Esta partición no fue utilizada durante el entrenamiento, lo que permitió obtener una estimación objetiva del desempeño de los modelos frente a datos no vistos.

Durante esta fase, los modelos generaron predicciones sobre las calificaciones finales de los estudiantes, las cuales fueron comparadas con los valores reales utilizando las métricas definidas.

Adicionalmente, se evaluó el funcionamiento del sistema de recomendaciones personalizadas. Para ello, se analizaron escenarios con perfiles estudiantiles específicos (por ejemplo, baja dedicación al estudio, alta exposición a redes sociales o escasa participación en tutorías), verificando que las sugerencias generadas fueran coherentes, relevantes y potencialmente útiles desde un enfoque académico y práctico.

4.8.3 Comparación de Modelos

Se realizó una comparación sistemática entre los dos modelos evaluados: **Random Forest** y **Gradient Boosting**. Para cada uno, se recopilaron métricas tanto en la fase de entrenamiento como en la fase de validación, lo que permitió evaluar su capacidad de generalización y rendimiento predictivo.

A partir de la evaluación cuantitativa de los modelos de regresión implementados, se observaron resultados positivos en términos de precisión predictiva. El modelo **Random**

Forest mostró un buen desempeño general, con errores de predicción bajos y una capacidad razonable para explicar la variabilidad en las calificaciones estudiantiles.

Por otro lado, el modelo de **Gradient Boosting** logró superar ligeramente a Random Forest en la mayoría de las métricas consideradas, evidenciando una mejor capacidad de generalización y mayor ajuste a los patrones presentes en los datos. Esta mejora fue especialmente notoria en la reducción del error cuadrático medio y el incremento en la proporción de varianza explicada.

Ambos enfoques demostraron ser efectivos para modelar y predecir el rendimiento académico, capturando relaciones significativas entre las variables predictoras y la calificación final. Los resultados obtenidos brindan una base sólida para el análisis comparativo desarrollado en el capítulo de resultados, donde se profundiza en el comportamiento de cada modelo y se identifican sus fortalezas y áreas de mejora.

Capítulo V

Resultados

En este capítulo se presentan los resultados obtenidos del proceso de entrenamiento y evaluación de los modelos de predicción del rendimiento académico. Se incluyen tanto los modelos base (Random Forest y Gradient Boosting) como los ajustes realizados sobre ellos. Comenzamos presentando un resumen del conjunto de datos utilizado, incluyendo el número total de registros, la distribución de las calificaciones, y las técnicas de preprocesamiento aplicadas. Este contexto es fundamental para interpretar los resultados obtenidos en términos de precisión y generalización de los modelos.

5.1 Resumen del conjunto de datos

Como se indicó en el Capítulo IV, para esta investigación se trabajó con un conjunto de datos compuesto por **1,194 registros** correspondientes a estudiantes de Ingeniería de Software. El dataset contiene variables personales, académicas y contextuales, tales como edad, semestre, horas de estudio, uso de redes sociales, asistencia, entre otras, además de la **calificación final del semestre**, que constituye la variable objetivo del modelo.

5.1.1 Distribución de la variable objetivo

En cuanto a la distribución de la variable objetivo (nota final del semestre), se identificaron valores continuos en el rango de 0 a 20.

5.1.2 División del conjunto de datos

El dataset fue dividido en dos subconjuntos principales: **conjunto de entrenamiento** y **conjunto de prueba**, siguiendo el estándar de partición 80/20. El conjunto de entrenamiento fue posteriormente dividido en subconjuntos de entrenamiento y validación interna durante el entrenamiento de los modelos.

La tabla 5.1 resume esta partición:

Tabla 5.1: División del Dataset para Predicción Académica

Dataset Split	**Cantidad de muestras**
Conjunto de Entrenamiento	955
Conjunto de Prueba	239
Total	1,194

Durante el entrenamiento, el conjunto de entrenamiento fue subdividido nuevamente para realizar validación cruzada. En este caso, el 80% fue utilizado para entrenar los modelos y el 20% restante para validarlos durante el proceso de ajuste de hiperparámetros. La tabla 5.2 detalla esta segunda partición.

Tabla 5.2: División del Conjunto de Entrenamiento para Validación Interna

Dataset Split	**Cantidad de muestras**
Subconjunto de Entrenamiento	764
Subconjunto de Validación	191

5.1.3 Resumen de las particiones realizadas

A continuación, se presenta un resumen global de todas las divisiones realizadas sobre el conjunto de datos para las fases de entrenamiento, validación y prueba.

Tabla 5.3: Resumen del Dataset - Tarea de Predicción Académica

División	**Cantidad**	**Porcentaje**
Total (Dataset)	1,194	100%
Entrenamiento (ET)	955	80% del Dataset
Prueba (PR)	239	20% del Dataset
Entrenamiento Modelos	764	80% de ET
Validación Modelos	191	20% de ET

5.2 Resultados Cuantitativos

En esta sección se presentan los resultados obtenidos por los modelos propuestos, evaluados mediante las métricas especificadas en el Capítulo IV.

Los modelos fueron implementados utilizando Python y bibliotecas como Scikit-learn. Las métricas utilizadas incluyen: MAE, RMSE, MedAE, SMAPE y el coeficiente de determinación R^2 para la regresión.

5.2.1 Resultados cuantitativos para la tarea de regresión académica

Métricas de evaluación

Los modelos fueron entrenados con el 80% del conjunto de datos y evaluados con el 20% restante. A continuación, se presentan los resultados cuantitativos obtenidos por los dos modelos de regresión utilizados: *Random Forest Regressor* y *Gradient Boosting Regressor*.

Tabla 5.4: Comparación de rendimiento durante el entrenamiento de los modelos en el conjunto de datos de entrenamiento para la tarea de regresión académica

Modelo	**MAE**	**RMSE**	**MedAE**	**SMAPE (%)**	**R^2**
Random Forest	1.60	2.00	1.42	14.38%	0.77
Gradient Boosting	**1.52**	**1.84**	**1.37**	17.01%	**0.81**

Como se observa en la Tabla 5.4, el modelo **Gradient Boosting Regressor** logró un mejor desempeño general en la mayoría de las métricas, destacando un mayor coeficiente de determinación (R^2) y menor error cuadrático medio.

5.3 Resultados de la Generación de Recomendaciones Personalizadas

Evaluación del sistema de recomendaciones

Para evaluar la utilidad del sistema de recomendaciones personalizadas, se simularon distintos perfiles estudiantiles con características frecuentemente asociadas a un bajo rendimiento académico (por ejemplo, escasa dedicación al estudio, alta exposición a redes sociales o nula participación en tutorías).

El sistema genera sugerencias mediante simulaciones inteligentes basadas en el modelo *Gradient Boosting Regressor*. A partir del perfil ingresado, se modifican virtualmente variables clave (como las horas de estudio o el nivel de asistencia), y se estima cuantitativamente el impacto que tendría cada cambio sobre la calificación final.

Esto permite ofrecer recomendaciones personalizadas, fundamentadas en datos y orientadas a la acción.

Lógica del sistema de recomendaciones

El motor de recomendaciones analiza el impacto potencial de las siguientes acciones:

- Aumentar en una hora diaria el tiempo de estudio si actualmente es menor a 5 horas.
- Incrementar la asistencia promedio al 90% si se encuentra por debajo de ese valor.
- Reducir el uso de redes sociales a un máximo de 2 horas diarias si supera las 4 horas.
- Dormir al menos 7 horas por día.
- Realizar al menos 3 horas semanales de ejercicio físico.
- Estudiar al menos 5 días a la semana.
- Dedicarse al desarrollo de habilidades específicas por al menos 3 horas diarias.
- Asistir a sesiones de tutoría si actualmente no se participa en ellas.

Cada recomendación se evalúa en función de la mejora estimada en la calificación. Solo se presentan al usuario aquellas sugerencias cuya mejora proyectada supera un umbral mínimo (por ejemplo, +0.05 puntos).

A continuación, se presentan el resultado real de recomendaciones generadas por el sistema para distintos perfiles académicos simulados:

Tabla 5.5: Ejemplos de recomendaciones personalizadas simuladas

Perfil Estudiantil Simulado	**Recomendaciones Generadas (con mejora estimada)**
4h de estudio/día, 80% de asistencia, 5h en redes sociales	Estudiar +1h al día (+0.45 pts) Subir asistencia al 90% (+0.30 pts) Reducir redes sociales a 2h (+0.41 pts)
2h de sueño, sin tutorías, 1h de ejercicio semanal	Dormir al menos 7h (+0.52 pts) Asistir a tutorías docentes (+0.38 pts) Realizar al menos 3h de ejercicio (+0.25 pts)
Estudia solo 2 días/semana, 0h en habilidades	Estudiar al menos 5 días a la semana (+0.36 pts) Practicar habilidades 3h al día (+0.48 pts)

Capítulo VI

Discusión de los Resultados

6.1 Análisis General

Los modelos desarrollados —*Random Forest Regressor* y *Gradient Boosting Regressor*— fueron entrenados y evaluados utilizando un conjunto de datos con variables relacionadas al estilo de vida y desempeño académico de estudiantes universitarios. La evaluación se llevó a cabo sobre un conjunto de prueba previamente reservado (20% del total), lo cual permitió obtener una estimación objetiva del rendimiento de cada modelo frente a datos no vistos.

En líneas generales, ambos modelos mostraron un desempeño satisfactorio, siendo capaces de predecir la nota final del semestre con errores relativamente bajos y un buen nivel de explicación de la varianza.

6.2 Análisis de Métricas de Evaluación

Los resultados muestran que el modelo **Gradient Boosting** superó al **Random Forest** en todas las métricas evaluadas. El **MAE** y **RMSE** más bajos indican que sus predicciones fueron más precisas y con menor dispersión respecto a los valores reales. La mediana del error absoluto (**MedAE**) también fue inferior, lo que sugiere que el modelo comete errores menores en la mayoría de los casos.

En términos relativos, el **SMAPE** —que expresa el error porcentual medio— también fue más bajo en el modelo de *Gradient Boosting*, con un 8.67%, en comparación con el 9.84% de *Random Forest*. Este resultado es particularmente importante en contextos donde se desea interpretar el error en función del valor predicho.

Finalmente, el coeficiente de determinación ($\mathbf{R^2}$) fue mayor para el modelo *Gradient Boosting* (0.88), lo que implica que es capaz de explicar un 88% de la variabilidad en las notas finales de los estudiantes. Este es un indicador fuerte del buen ajuste del modelo sin evidencia significativa de sobreajuste, ya que se calculó sobre el conjunto de prueba.

Tabla 6.1: Comparación de métricas entre modelos

Métrica	**Random Forest**	**Gradient Boosting**
MAE	1.18	1.06
RMSE	1.47	1.32
MedAE	1.03	0.94
SMAPE (%)	9.84	8.67
R^2	0.84	0.88

Capítulo VII

Conclusiones y Recomendaciones

7.1 Conclusiones

Como primera conclusión, se evidenció que los resultados obtenidos en los experimentos de predicción respaldan la efectividad de los modelos de aprendizaje automático, específicamente **Random Forest Regressor** y **Gradient Boosting Regressor**, en la tarea de estimar el rendimiento académico de los estudiantes universitarios. Ambos modelos demostraron un desempeño robusto al presentar bajos valores en métricas de error como el *Mean Absolute Error* (MAE), el *Root Mean Squared Error* (RMSE) y el *Median Absolute Error* (MedAE), así como un alto coeficiente de determinación (R^2), lo que indica una adecuada capacidad de generalización del comportamiento académico a partir de los datos de entrada.

En particular, el modelo **Gradient Boosting Regressor** mostró un mejor equilibrio entre complejidad y precisión, superando al modelo Random Forest en la mayoría de las métricas evaluadas. Este modelo alcanzó un coeficiente de determinación de $R^2 = 0.88$,

lo que sugiere que puede explicar el 88% de la varianza en las calificaciones finales, convirtiéndolo en una herramienta sólida para la predicción educativa.

Como segunda conclusión, se demostró que el uso de variables académicas, personales y contextuales —tales como la asistencia promedio, las horas de estudio diario, el uso de redes sociales, la participación en tutorías, las horas de sueño y el ejercicio físico semanal— permite construir un modelo más integral y representativo del rendimiento académico real de los estudiantes. Esta combinación de variables heterogéneas resultó clave para la calidad predictiva alcanzada por los modelos.

Asimismo, se evidenció que el uso de métricas de regresión como MAE y RMSE resultó apropiado para evaluar el rendimiento del sistema, ya que permiten cuantificar de forma clara el margen de error esperado entre la nota real y la nota predicha. Estas métricas ofrecieron una visión más precisa y continua del desempeño del modelo frente a la realidad de los datos.

En resumen, los resultados experimentales validan la eficacia de los modelos de regresión Random Forest y Gradient Boosting como herramientas predictivas confiables para apoyar la toma de decisiones en contextos educativos. Su aplicación puede resultar valiosa en programas de tutoría, sistemas de alerta temprana o generación de recomendaciones personalizadas orientadas a mejorar el desempeño académico de los estudiantes.

7.2 Recomendaciones

Con base en la ejecución de los experimentos y el análisis de resultados, se proponen las siguientes recomendaciones:

- **Ampliar el tamaño y diversidad del dataset:** Se recomienda recopilar más datos

de diferentes ciclos, carreras y contextos académicos para mejorar la capacidad de generalización del modelo. Un conjunto de datos más amplio permitirá captar patrones más representativos y robustos.

- **Incluir variables longitudinales:** Incorporar información histórica del estudiante (como evolución de notas, asistencia o tutorías a lo largo de varios semestres) podría mejorar la capacidad del modelo para anticipar comportamientos futuros y detectar tendencias académicas.

- **Evaluar la utilidad de nuevas variables psicoeducativas:** Se sugiere explorar la inclusión de variables relacionadas con la motivación, el estrés, la salud mental o el entorno familiar, ya que pueden tener un impacto significativo en el desempeño académico.

- **Implementar el sistema de recomendaciones en entornos reales:** La integración de la plataforma predictiva en el sistema de información académica de la facultad permitiría probar su efectividad en la práctica y ofrecer retroalimentación útil tanto para estudiantes como para docentes y asesores.

- **Monitoreo continuo del rendimiento del modelo:** Es importante evaluar periódicamente el comportamiento de los modelos predictivos y ajustarlos conforme se disponga de nuevos datos o cambien las condiciones académicas. El mantenimiento del modelo es crucial para conservar su precisión a lo largo del tiempo.

- **Complementar con modelos explicables (Explainable AI):** Para facilitar la comprensión por parte de autoridades académicas, estudiantes y padres, se recomienda incorporar modelos o herramientas que expliquen las decisiones del modelo (por

ejemplo, SHAP o LIME), ayudando a interpretar por qué un estudiante ha sido clasificado en determinada categoría.

Finalmente, se recomienda seguir explorando otras técnicas de aprendizaje automático y estrategias híbridas que combinen modelos estadísticos con enfoques de inteligencia artificial para fortalecer los sistemas de apoyo educativo en entornos universitarios.

Referencias

Ahmed, E. (2024). Student performance prediction using machine learning algorithms. *Applied Computational Intelligence and Soft Computing*.

Albreiki, B., Zaki, N., and Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Education Sciences*, 11(9):552.

Alsubaie, M. N. (2023). Predicting student performance using machine learning to enhance the quality assurance of online training via maharat platform. *Alexandria Engineering Journal*, 69:323–339.

Alsubhi, B., Alharbi, B., Aljojo, N., Banjar, A., Tashkandi, A., Alghoson, A., and Al-Tirawi, A. (2023). Effective feature prediction models for student performance. *Engineering, Technology & Applied Science Research*, 13(5):11937–11944.

Atalla, S., Daradkeh, M., Gawanmeh, A., Khalil, H., Mansoor, W., Miniaoui, S., and Himeur, Y. (2023). An intelligent recommendation system for automating academic advising based on curriculum analysis and performance modeling. *Mathematics*, 11(5):1098.

Balabied, S. A. A. and Eid, H. F. (2023). Utilizing random forest algorithm for early detection of academic underperformance in open learning environments. *PeerJ Computer Science*, 9:e1708.

Buenaño-Fernández, D., Gil, D., and Luján-Mora, S. (2019). Application of machine learning in predicting performance for computer engineering students: A case study. *Sustainability*, 11(10):2833.

Chen, M. and Liu, Z. (2024). Predicting performance of students by optimizing tree components of random forest using genetic algorithm. *Heliyon*, 10(12).

Delahoz-Dominguez, E. and Hijón-Neira, R. (2024). Recommender system for university degree selection: A socioeconomic and standardised test data approach. *Applied Sciences*, 14(18):8311.

Guanin-Fajardo, J. H., Guaña-Moya, J., and Casillas, J. (2024). Predicting academic success of college students using machine learning techniques. *Data*, 9(4):60.

Hashim, A. S., Awadh, W. A., and Hamoud, A. K. (2020). Student performance prediction model based on supervised machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering*, volume 928, page 032019. IOP Publishing.

Jiao, P., Ouyang, F., Zhang, Q., and Alavi, A. H. (2022). Artificial intelligence-enabled prediction model of student academic performance in online engineering education. *Artificial Intelligence Review*, 55(8):6321–6344.

Nachouki, M., Mohamed, E. A., Mehdi, R., and Abou Naaj, M. (2023). Student course grade prediction using the random forest algorithm: Analysis of predictors' importance. *Trends in Neuroscience and Education*, 33:100214.

Namoun, A. and Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1):237.

Nizar, J., Sharmila, R., and Jaseena, K. U. (2024). A random forest model for prediction of software engineering skill set among computer science students through explainable ai. *International Journal of Intelligent Systems and Applications in Engineering (IJISAE)*, 12(21s).

Oyedeji, A. O., Salami, A. M., Folorunsho, O., and Abolade, O. R. (2020). Analysis and prediction of student academic performance using machine learning. *JITCE (Journal of Information Technology and Computer Engineering)*, 4(1):10–15.

Pan, S. and Dai, W. (2024). Research on student performance prediction based on random forest algorithm. In *Proceedings of the 2024 International Symposium on Artificial Intelligence for Education (ISAIE '24)*, pages 511–515. Association for Computing Machinery.

Sekeroglu, B., Abiyev, R., Ilhan, A., Arslan, M., and Idoko, J. B. (2021). Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies. *Applied Sciences*, 11(22):10907.

Selvakumari, S. et al. (2023). Design of a prediction model to predict students' performance using educational data mining and machine learning. *Engineering Proceedings*, 59(1):25.

Wang, J. and Yu, Y. (2025). Machine learning approach to student performance prediction of online learning. *PLOS ONE*.

Yang, Y., Zhao, S., An, S., Li, X., and Zhang, Y. (2025). Student academic performance prediction via hypergraph and tabnet. *Journal of Big Data*, 12(1):119.

Zhao, L., Ren, J., Zhang, L., and Zhao, H. (2023). Quantitative analysis and prediction of academic performance of students using machine learning. *Sustainability*, 15(16):12531.

Zhu, L., Zhang, S., Wei, Y., Tu, X., Huang, Y., and Wu, M. (2025). Predicting student performance in software engineering education using random forest: A data-driven approach based on subjective assessments. In *Proceedings of the 2025 International Conference on Digital Education and Information Technology*, pages 136–141. Association for Computing Machinery.