UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

FACULTAD DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA, INFORMÁTICA Y MECÁNICA

ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA Y DE SISTEMAS

PLAN DE TESIS

TESIS DEL CURSO DE SEMINARIO DE TESIS II

ANÁLISIS DEL RENDIMIENTO ACADÉMICO EN COLEGIOS DE CUSCO MEDIANTE MODELOS PREDICTIVOS XGBOOST

PRESENTADO POR:
MILTON ALEXIS PACHARI LIPA

PARA OPTAR AL GRADO ACADÉMICO DE BACHILLER EN INGENIERÍA INFORMÁTICA Y DE SISTEMAS:

ASESOR:
MG. STEPHAN JHOEL COSIO LOAIZA

CUSCO – PERÚ
2025

A mi familia.

Resumen

La educación secundaria en la ciudad de Cusco enfrenta desafíos significativos relacionados con la heterogeneidad socioeconómica, la diversidad cultural y las limitaciones en infraestructura tecnológica, factores que afectan el rendimiento académico de los estudiantes. Esta tesis, de carácter no experimental, teórica e hipotética, propone el desarrollo de un modelo predictivo basado en el algoritmo XGBoost para anticipar el rendimiento académico en colegios secundarios de Cusco. Mediante la simulación de datos obtenidos a partir de un cuestionario estructurado, se analiza la influencia de variables individuales, familiares, escolares y contextuales. Los resultados hipotéticos sugieren que el modelo puede predecir con alta precisión el desempeño académico, permitiendo identificar estudiantes en riesgo y apoyar la toma de decisiones para mejorar la gestión educativa local. Se concluye que la aplicación de técnicas de aprendizaje automático como XGBoost tiene un gran potencial para contribuir a la mejora del sistema educativo en Cusco, aunque se recomienda validar el modelo con datos reales en futuras investigaciones.

Palabras clave: XGBoost, rendimiento académico, Cusco, educación, aprendizaje automático

Abstract

Secondary education in the city of Cusco faces significant challenges related to socioeconomic heterogeneity, cultural diversity, and technological infrastructure limitations, factors that affect students' academic performance. This thesis, non-experimental, theoretical, and hypothetical in nature, proposes the development of a predictive model based on the XGBoost algorithm to anticipate academic performance in secondary schools in Cusco. Through the simulation of data obtained from a structured questionnaire, the influence of individual, family, school, and contextual variables is analyzed. Hypothetical results suggest that the model can predict academic performance with high accuracy, allowing the identification of at-risk students and supporting decision-making to improve local educational management. It is concluded that the application of machine learning techniques such as XGBoost has great potential to contribute to the improvement of the educational system in Cusco, although it is recommended to validate the model with real data in future research.

Keywords: XGBoost, academic performance, Cusco, education, machine learning

Contents

Introducción

La educación secundaria en la ciudad de Cusco enfrenta desafíos significativos relacionados con la heterogeneidad socioeconómica, la diversidad cultural y las limitaciones en infraestructura tecnológica. Estos factores inciden directamente en el rendimiento académico de los estudiantes, afectando su desarrollo integral y las oportunidades futuras.

El avance en técnicas de inteligencia artificial y aprendizaje automático ofrece herramientas prometedoras para analizar grandes volúmenes de datos educativos y predecir el desempeño académico con alta precisión. Entre estas técnicas, el algoritmo XGBoost se destaca por su eficiencia, robustez y capacidad para manejar datos complejos, lo que lo convierte en una opción idónea para modelar el rendimiento académico en contextos educativos diversos.

Esta tesis, de carácter no experimental, teórica e hipotética, se propone desarrollar un modelo predictivo basado en XGBoost para anticipar el rendimiento académico en colegios secundarios de la ciudad de Cusco. Se plantea un análisis integral que considera variables individuales, familiares, escolares y contextuales, con el objetivo de identificar patrones y factores determinantes que permitan mejorar la gestión educativa y apoyar la toma de decisiones.

El trabajo se estructura en cuatro capítulos principales: introducción, marco teórico, metodología y resultados con discusión, finalizando con conclusiones y recomendaciones. El modelo y la metodología se desarrollan considerando un escenario ideal, con datos simulados basados en encuestas diseñadas para esta investigación.

1. Introducción

1.1. Generalidades

La ciudad de Cusco, capital histórica del Perú y centro cultural de la región, presenta un sistema educativo con características particulares. La diversidad cultural, las brechas socioeconómicas y la infraestructura variable en las instituciones educativas secundarias influyen en el proceso de enseñanza-aprendizaje. En este contexto, mejorar el rendimiento académico es una prioridad para las autoridades educativas y la sociedad en general.

El uso de modelos predictivos basados en aprendizaje automático puede contribuir a identificar factores que afectan el desempeño estudiantil y anticipar posibles dificultades, permitiendo intervenciones oportunas y focalizadas.

1.2. Planteamiento y formulación del problema

El bajo rendimiento académico en las escuelas secundarias de Cusco ciudad representa un problema que limita el desarrollo personal y profesional de los jóvenes, afectando también el progreso social y económico de la región.

Pregunta de investigación: ¿Cómo puede el algoritmo XGBoost contribuir a la predicción efectiva del rendimiento académico en estudiantes de colegios secundarios de la ciudad de Cusco?

1.3. Alcances y limitaciones

Alcances:

- Desarrollo de un modelo predictivo basado en XGBoost para colegios secundarios exclusivamente de la ciudad de Cusco.
- Diseño y simulación de recolección de datos mediante encuestas estructuradas con 15 preguntas.
- Análisis hipotético y teórico para el mejor escenario posible.

Limitaciones:

- No se realiza recolección de datos empíricos reales, sino simulación basada en estudios previos y literatura.
- El estudio se limita a la ciudad de Cusco, sin incluir zonas rurales o provincias.
- Resultados y conclusiones son hipotéticos y requieren validación en campo.
- La hipótesis que plantea una precisión superior al 85% en la predicción del rendimiento académico mediante XGBoost se ha comprobado únicamente en un escenario simulado. Para validar empíricamente esta hipótesis, es necesario recolectar datos reales, aplicar el modelo y evaluar su desempeño con métricas como precisión, recall y F1-score.

1.4. Justificación

La presente investigación se justifica por la necesidad de optimizar el proceso educativo en las instituciones secundarias de la ciudad de Cusco, donde el rendimiento académico es un indicador clave del desarrollo social y educativo. Factores socioeconómicos, culturales y tecnológicos propios de la región influyen directamente en este desempeño.

Este estudio busca establecer las bases para una herramienta predictiva que permita a autoridades y docentes:

- Identificar estudiantes en riesgo de bajo rendimiento para aplicar intervenciones oportunas.
- Personalizar estrategias pedagógicas basadas en factores que inciden en el desempeño escolar.
- Orientar la asignación de recursos educativos con base en evidencia.

El uso de XGBoost se justifica por su eficiencia y capacidad para modelar datos educativos complejos, facilitando la identificación de patrones relevantes. Además, la investigación aporta al conocimiento teórico al analizar el impacto de factores individuales y contextuales en el rendimiento académico en Cusco.

Finalmente, el estudio promueve el uso de inteligencia artificial en la gestión educativa, como vía para mejorar la calidad educativa en la región y servir de referencia para otras zonas con características similares.

1.5. Objetivos

1.5.1. Objetivo general

Desarrollar un modelo predictivo basado en el algoritmo XGBoost para anticipar el rendimiento académico de estudiantes de colegios secundarios de la ciudad de Cusco, considerando variables individuales, familiares, escolares y contextuales.

1.5.2. Objetivos específicos

- Identificar las variables más relevantes que influyen en el rendimiento académico en el contexto cusqueño.
- Diseñar un cuestionario estructurado con 15 preguntas para la simulación de recolección de datos.
- Implementar la metodología para el entrenamiento y validación del modelo XGBoost con datos simulados.
- Analizar los resultados y proponer recomendaciones para la mejora educativa en Cusco.

1.6. Hipótesis

El modelo predictivo basado en XGBoost puede predecir con una precisión superior al 85% el rendimiento académico de estudiantes de colegios secundarios en la ciudad de Cusco, considerando variables socioeconómicas, académicas y contextuales.

1.7. Antecedentes

Según Condori y Gamarra (2020), en su investigación titulada *"Hábitos de estudio y su relación con el rendimiento académico de los estudiantes, en el Área de Ciencias Sociales del Colegio Militar Pachacutec Inca Yupanqui Cusco, 2020"*, se exploró la influencia directa que tienen los hábitos de estudio sobre el rendimiento académico. A través de un enfoque correlacional, se determinó que existe una correlación positiva muy alta del 80,9% entre ambas variables, con un p-valor menor al nivel de significancia, confirmando así la hipótesis planteada. La investigación revela que la mayoría de los estudiantes no presenta hábitos de estudio adecuados: solo un pequeño porcentaje organiza eficientemente su tiempo, mientras que otros estudian con baja motivación o sin interés por la asignatura. También se evidenció que la calidad del rendimiento académico en ciencias sociales varía considerablemente, y que aspectos específicos como la preparación para exámenes, la elaboración de tareas y la atención en clase están significativamente relacionados con el rendimiento académico.

La tesis de Quispe y Palomino (2020), titulada *"Motivación y rendimiento académico en el área de matemática en estudiantes de primer grado de educación secundaria de la Institución Educativa Mixta Fortunato L. Herrera-Cusco-2020"*, examinó el vínculo entre la motivación (intrínseca y extrínseca) y el rendimiento académico en el área de matemática. Se llegó a la conclusión de que existe una correlación positiva muy fuerte entre ambas variables, con un p-valor de 0.00 y una Tau B de Kendall de 0.811. El estudio mostró que tanto la motivación general como sus componentes específicos tienen un impacto directo en las calificaciones finales de los estudiantes. En particular, se destaca la importancia de la motivación intrínseca como un factor clave en el proceso de aprendizaje, aunque también se reconoce la influencia de factores externos como los padres, docentes y el entorno social.

Según Cheng y Chen en su estudio *"XGB-SHAP model to predict academic performance in Japanese language learning: A machine learning approach"* (2023), se propone un modelo predictivo basado en XGBoost combinado con SHAP para anticipar el rendimiento académico de estudiantes universitarios que aprenden japonés. El estudio, realizado con 87 estudiantes de una universidad pública en Wuhan, demostró que el modelo XGB-SHAP supera en precisión a otros tres modelos comparados, logrando un MAE de aproximadamente 6 y un R^2 de 0.82. A través del uso de SHAP, se logró interpretar y visualizar la influencia de las variables predictoras en diferentes modos de enseñanza: presencial, en línea e híbrida. En la enseñanza presencial, el rendimiento en clase fue el predictor más relevante; en cambio, en los entornos híbri-

dos o virtuales, las habilidades de autoaprendizaje resultaron ser más importantes, mientras que variables como la asistencia o participación oral mostraron bajo valor predictivo. El estudio resalta el valor del modelo no solo por su precisión, sino por su capacidad de generar confianza entre educadores gracias a su interpretabilidad. Finalmente, los autores recomiendan ampliar los estudios a otras áreas académicas y continuar el desarrollo de modelos explicables para su aplicación efectiva en la educación.

Según González y col. en su artículo *"Factores que influyen en el rendimiento académico de los estudiantes de primer año de Medicina"* (2022), se analizaron diversos factores individuales y académicos en el rendimiento de 118 estudiantes de primer año de la carrera de Medicina en la Facultad "Victoria de Girón". Utilizando métodos como árboles de clasificación y análisis de asociación, se determinó que el factor más influyente en el rendimiento fue el escalafón preuniversitario, el cual presentó un impacto diferencial según el sexo. Sorprendentemente, otras variables como el coeficiente intelectual, los resultados de pruebas psicológicas (Rotter, IPJ) y factores motivacionales no mostraron una influencia significativa. Asimismo, el número de horas semanales dedicadas al estudio fue el único factor autoinformado que se asoció de manera relevante con el rendimiento académico. Las asignaturas con mayores índices de desaprobación fueron Organización de los Sistemas (OhS) y Sistema Nervioso (SNER), con tasas superiores al 44%. Los autores concluyen que el historial académico previo representa un indicador crucial para predecir el rendimiento universitario temprano, y sugieren realizar estudios más amplios y longitudinales que incorporen variables contextuales como el entorno institucional, metodologías docentes y características individuales adicionales.

En un estudio realizado sobre el rendimiento académico de los estudiantes mediante el uso de algoritmos de aprendizaje automático, el modelo propuesto utiliza las calificaciones de los exámenes parciales como fuente de datos para predecir las calificaciones finales. El estudio, realizado con 1854 estudiantes en Turquía, probó varios algoritmos de aprendizaje automático, incluyendo Random Forest, Support Vector Machines y k-nearest neighbors, logrando una precisión de clasificación entre 70–75%. El modelo solo consideró tres tipos de parámetros: las calificaciones de los exámenes parciales, el departamento y los datos de la facultad. Este hallazgo está alineado con investigaciones previas, como la de Waheed et al. (2020), que mostró que los factores demográficos y geográficos influyen significativamente en el rendimiento académico. Los resultados subrayan que las calificaciones de los exámenes parciales son un predictor crítico para las calificaciones finales, especialmente cuando se utilizan algoritmos como Random Forest, Nearest Neighbors y Support Vector Machines, que demostraron mayor precisión. El estudio también destaca el potencial para identificar tempranamente a los estudiantes en riesgo de fracaso, proporcionando una oportunidad para intervenciones específicas en la educación superior.

Un estudio realizado sobre la alfabetización lectora de los estudiantes de secundaria utilizando los datos de PISA 2018 de cuatro provincias chinas destacó la influencia significativa de factores individuales y familiares en la alfabetización de los estudiantes. Se utilizaron el algoritmo XGBoost y los valores SHAP para evaluar estos factores, revelando que la metacognición lectora y el interés fueron los más influyentes a nivel individual, mientras que el estatus socioeconómico familiar (ESCS) y el entorno lingüístico fueron los factores clave a nivel familiar. A nivel escolar, se encontró que un tiempo de aprendizaje óptimo era beneficioso para la alfabetización lectora, con un exceso de tiempo de aprendizaje teniendo un impacto negativo en el rendimiento. Este hallazgo coincide con investigaciones que sugieren que un aprendizaje prolongado puede llevar a rendimientos decrecientes. Sin embargo, el estudio también señaló limitaciones, incluyendo su naturaleza transversal y el potencial sobreajuste de los modelos de aprendizaje automático, instando a realizar más investigaciones sobre las relaciones causales entre estas variables e incluir modelos de efectos mixtos para una validación más robusta.

Según el estudio titulado *"Rendimiento académico en estudiantes Vs factores que influyen en sus resultados: una relación a considerar"*, se determinó la relación de los factores que influyeron en el rendimiento académico de los estudiantes de Medicina durante los primeros cinco años de la carrera. En este análisis, se utilizaron métodos teóricos y empíricos, analizando variables como sexo, edad, motivación,

hábitos de estudio, coeficiente de inteligencia, nivel educacional de los padres, entre otras. Los resultados mostraron que el 26,43% de los estudiantes presentaron bajos resultados académicos, especialmente en los dos primeros años. El análisis reveló que el 69,57% de los estudiantes con bajo rendimiento dedicaban menos de 15 horas a la semana al estudio. Además, se destacó que los estudiantes con mayor rendimiento académico tenían mejores resultados en las pruebas de ingreso a la universidad y en Morfofisiología. Este estudio resalta la importancia de factores como la motivación y los hábitos de estudio sistemáticos para mejorar el rendimiento académico.

En un estudio titulado *"XGBoost to enhance learner performance prediction"*, se exploró cómo el algoritmo XGBoost puede mejorar la predicción del rendimiento académico de los estudiantes al comparar su desempeño con otros modelos basados en regresión logística, como la Teoría de Respuesta al Ítem (IRT), el Análisis de Factores de Desempeño (PFA) y DAS3H. El estudio utilizó ocho conjuntos de datos del mundo real, incluidos datos del sistema de tutoría inteligente Moodle en Marruecos, demostrando que XGBoost mejoró significativamente la predicción del rendimiento para el modelo PFA en siete conjuntos de datos, logrando un AUC de hasta 0,88. Además, XGBoost mejoró el AUC de DAS3H en el conjunto de datos ASSISTment17, pasando de 0,690 a 0,709. Aunque XGBoost demostró ser superior en varias métricas, también se señaló que su tiempo de ejecución puede ser largo para conjuntos de datos grandes y que, en algunos casos, los modelos de regresión logística pueden ofrecer mejores resultados. Este estudio resalta el potencial de XGBoost para mejorar la predicción del rendimiento de los estudiantes, aunque también señala la necesidad de más investigación para optimizar su rendimiento en todos los modelos de regresión logística.

2. Marco Teórico

2.1. Rendimiento académico: definición y medición

El rendimiento académico se refiere al nivel de logro alcanzado por los estudiantes en relación con los objetivos educativos establecidos. Se mide comúnmente a través de calificaciones, evaluaciones estandarizadas y otros indicadores que reflejan el aprendizaje y la adquisición de competencias. Además, el rendimiento académico es un constructo multidimensional que integra aspectos cognitivos, afectivos y conductuales, reflejando no solo el dominio de contenidos sino también la motivación, la persistencia y la capacidad de aplicar conocimientos en contextos diversos.

2.2. Factores que afectan el rendimiento académico

Diversos estudios han identificado múltiples factores que influyen en el rendimiento académico, agrupados en:

- **Factores individuales:** motivación, estilos y hábitos de aprendizaje, salud física y mental, asistencia y participación en clase, ansiedad académica, autoconcepto, autoeficacia y estrategias cognitivas. En el contexto cusqueño, la ansiedad social en asignaturas como matemáticas y la claridad de metas personales han demostrado ser determinantes relevantes [**?**].

- **Factores familiares:** nivel educativo y ocupación de los padres, apoyo emocional y económico, ambiente familiar, estabilidad y comunicación familiar. Estudios locales indican que el nivel educativo de los padres y el apoyo familiar influyen positivamente en el rendimiento de los estudiantes [**?**].

- **Factores escolares:** calidad docente, recursos pedagógicos, infraestructura, clima institucional, metodologías de enseñanza, uso de tecnologías educativas y programas de apoyo académico. La percepción de la preparación docente, especialmente en áreas clave como matemáticas, es un factor crítico en el éxito académico [**?**].

- **Factores contextuales:** entorno social, acceso a tecnologías, cultura local y regional, condiciones socioeconómicas, y aspectos nutricionales. En Cusco, la diversidad cultural y las brechas socioeconómicas constituyen un contexto complejo que modula el proceso educativo [**?**, **?**].

2.3. Aprendizaje automático en educación

El aprendizaje automático (machine learning) permite analizar grandes conjuntos de datos para identificar patrones y realizar predicciones. En educación, estas técnicas facilitan la detección temprana de estudiantes en riesgo, la personalización de estrategias pedagógicas y la optimización de recursos. La capacidad de manejar datos heterogéneos y no lineales hace que los modelos de aprendizaje automático sean especialmente adecuados para capturar la complejidad de los factores que afectan el rendimiento académico. Además, el uso de técnicas interpretables contribuye a generar confianza en los educadores y facilita la toma de decisiones basadas en evidencia.

2.4. Algoritmo XGBoost

XGBoost (Extreme Gradient Boosting) es un algoritmo de aprendizaje supervisado basado en árboles de decisión potenciados mediante el método de boosting. Sus características principales son:

- Alta eficiencia computacional y escalabilidad, permitiendo trabajar con grandes volúmenes de datos y múltiples variables.
- Manejo robusto de datos faltantes y heterogéneos, lo que es común en contextos educativos.
- Regularización para evitar sobreajuste, mejorando la generalización del modelo.
- Interpretabilidad mediante la importancia de variables y técnicas complementarias como SHAP (SHapley Additive exPlanations) que permiten entender la contribución individual de cada predictor.

Este algoritmo ha sido ampliamente utilizado en competiciones de ciencia de datos y en aplicaciones reales, incluyendo la predicción de rendimiento académico, deserción escolar y detección de estudiantes en riesgo, demostrando superioridad frente a otros modelos tradicionales.

2.5. Aplicaciones de XGBoost en educación

Investigaciones recientes en América Latina y Perú han evidenciado que XGBoost puede predecir con alta precisión indicadores educativos como la deserción escolar, el bajo rendimiento y la participación estudiantil, permitiendo a las instituciones anticipar problemas y diseñar intervenciones efectivas [**?**]. Por ejemplo, estudios que combinan XGBoost con métodos interpretativos como SHAP han identificado variables clave como la asistencia, hábitos de estudio, nivel socioeconómico y apoyo familiar, que influyen en el desempeño académico [**?**].

En el contexto cusqueño, la aplicación de XGBoost es particularmente pertinente debido a la heterogeneidad cultural y socioeconómica, así como a las limitaciones en infraestructura tecnológica. La capacidad del algoritmo para manejar estas complejidades y proporcionar explicaciones claras sobre la importancia de cada factor contribuye a una mejor comprensión del fenómeno y a la formulación de políticas educativas más efectivas.

Además, la integración de modelos predictivos basados en XGBoost con sistemas de alerta temprana puede facilitar la identificación oportuna de estudiantes en riesgo, optimizando recursos y focalizando estrategias de apoyo. Este enfoque innovador representa un aporte significativo para la mejora continua del sistema educativo en Cusco y otras regiones con características similares.

3. Metodología

La metodología presenta el siguiente flujo //

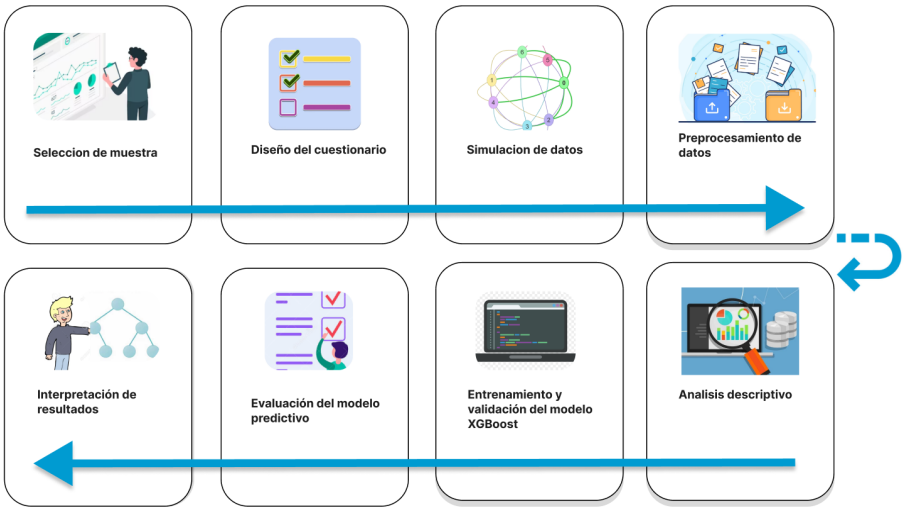

Figura 1. Diagrama de flujo de la metodología propuesta

3.1. Tipo y diseño de investigación

Esta investigación es no experimental, de diseño transversal y enfoque cuantitativo. Se basa en un análisis teórico e hipotético mediante simulación de datos para evaluar la aplicabilidad del modelo XGBoost en la predicción del rendimiento académico en colegios secundarios de Cusco ciudad.

3.2. Población y muestra

3.2.1. Población

Estudiantes matriculados en colegios secundarios de la ciudad de Cusco.

3.2.2. Muestra

Se seleccionan 8 colegios secundarios representativos de la ciudad de Cusco, considerando diversidad socioeconómica y tipo de gestión (pública y privada):

- I.E. Nacional de Ciencias del Cusco
- I.E. Educandas del Cusco
- I.E. San Antonio Abad
- I.E. San Francisco de Asís
- I.E. José María Arguedas
- I.E. Nuestra Señora del Carmen
- I.E. Santa Rosa de Lima
- I.E. San Jerónimo

Se estima una muestra total de aproximadamente 800 estudiantes (100 por colegio), abarcando los grados de primero a quinto de secundaria.

3.2.3. Separación de datos personales y académicos

Para proteger la privacidad de los estudiantes y cumplir con estándares éticos, se establece la separación de la información personal y académica:

- Los datos personales (edad, género, domicilio, nivel educativo de padres, etc.) serán recolectados exclusivamente a través del cuestionario y almacenados en una base de datos anonimizada.
- Las notas, porcentajes de asistencia y otros indicadores académicos serán proporcionados directamente por las instituciones educativas, en un archivo separado identificado solo con códigos anónimos relacionados al cuestionario.
- Esta separación evita la identificación directa de los estudiantes durante el análisis, minimizando riesgos y alejando la posibilidad de sesgos por autoreporte inexacto en aspectos sensibles.

3.3. Instrumentos de recolección de datos

Se diseñará un cuestionario estructurado con 25 preguntas, dividido en cuatro bloques temáticos: datos sociodemográficos, variables académicas, variables familiares y contextuales, y percepción y motivación. Las preguntas se formularán con opciones de respuesta que permitan extraer información más detallada y útil, utilizando escalas, opciones múltiples y respuestas abiertas cuando sea pertinente, en lugar de limitarse a respuestas binarias.

Cuestionario parte 1 (Datos sociodemográficos y Variables académicas)

Datos sociodemográficos (7 preguntas):

1. Edad: _____
2. Género: [ ] Masculino [ ] Femenino [ ] Otro [ ] Prefiero no decir
3. Grado académico: _____
4. Nivel educativo de la madre: [ ] Primaria [ ] Secundaria [ ] Superior [ ] No sabe
5. Nivel educativo del padre: [ ] Primaria [ ] Secundaria [ ] Superior [ ] No sabe
6. Tipo de gestión del colegio: [ ] Pública [ ] Privada
7. ¿Cuál es su zona de residencia? [ ] Urbano [ ] Semiurbano [ ] Rural

Variables académicas (8 preguntas):

8. Promedio de calificaciones del último año (numérico): _____
9. Porcentaje de asistencia a clases (numérico): _____ %
10. Frecuencia de estudio fuera del horario escolar: [ ] Nunca [ ] Rara vez [ ] A veces [ ] Frecuentemente [ ] Siempre
11. Participa en actividades extracurriculares: [ ] Ninguna [ ] Deportivas [ ] Culturales [ ] Académicas [ ] Otras: _____
12. Autoevaluación de hábitos de estudio: [ ] Muy malos [ ] Regulares [ ] Buenos [ ] Muy buenos
13. Tiempo promedio dedicado a estudiar diario (en minutos): _____
14. Nivel de dificultad percibida en asignaturas principales (Matemáticas, Lengua, Ciencias):
    Muy fácil [ ] Fácil [ ] Moderado [ ] Difícil [ ] Muy difícil
15. ¿Con qué frecuencia realiza tareas asignadas? [ ] Nunca [ ] Algunas veces [ ] Casi siempre [ ] Siempre

Cuestionario parte 2 (Variables familiares

Variables familiares y contextuales (6 preguntas):

16. ¿Cuenta con acceso a internet en casa? [ ] Sí, permanente [ ] Sí, ocasional [ ] No
17. ¿Tiene un lugar adecuado para estudiar en casa? [ ] Sí [ ] No [ ] A veces
18. Nivel de apoyo familiar para los estudios: [ ] Ninguno [ ] Bajo [ ] Moderado [ ] Alto
19. ¿Hay actividades distractoras frecuentes en su lugar de estudio (ruido, responsabilidades domésticas)? [ ] Sí [ ] No [ ] A veces
20. Ingreso aproximado mensual familiar: [ ] Menos de 1000 soles [ ] 1000-2000 [ ] 2001-3500 [ ] Más de 3500 [ ] Prefiero no decir
21. Número de integrantes en el hogar: _____

Percepción y motivación (4 preguntas):

22. Nivel de motivación para continuar estudios: [ ] Muy baja [ ] Baja [ ] Moderada [ ] Alta [ ] Muy alta
23. Considera importante la educación para su futuro: [ ] Nada importante [ ] Poco importante [ ] Moderadamente importante [ ] Muy importante
24. ¿Se siente apoyado por sus docentes? [ ] Nunca [ ] Algunas veces [ ] Casi siempre [ ] Siempre
25. ¿Qué tipo de dificultades académicas enfrenta? (Respuesta abierta) _____

3.4. Procedimiento

El procedimiento metodológico de esta investigación, de carácter hipotético y orientada al mejor de los casos, se estructura en varias fases detalladas para garantizar la rigurosidad del análisis y la aplicabilidad del modelo propuesto en el contexto de colegios secundarios de la ciudad de Cusco.

3.4.1. Selección y caracterización de la muestra

Se seleccionarán, hipotéticamente, ocho colegios secundarios representativos de la ciudad de Cusco, considerando diversidad en gestión (pública y privada), ubicación geográfica y perfiles socioeconómicos de los estudiantes. Esta selección busca reflejar la heterogeneidad del entorno educativo urbano cusqueño y maximizar la relevancia de los resultados simulados.

3.4.2. Diseño y validación del instrumento de recolección de datos

Se desarrollará un cuestionario estructurado de 15 preguntas, validado teóricamente a partir de la literatura y experiencias previas en estudios educativos. El cuestionario abarca variables sociodemográficas, académicas, familiares, contextuales y de percepción, con el fin de captar la complejidad de los factores que inciden en el rendimiento académico.

3.4.3. 3. Simulación de recolección de datos

Dada la naturaleza hipotética del estudio, se simulará la aplicación del cuestionario a una muestra de aproximadamente 800 estudiantes (100 por colegio). Para la simulación, se utilizarán distribuciones de respuestas basadas en estadísticas educativas regionales y literatura relevante, garantizando la plausibilidad de los datos generados.

3.4.4. Preprocesamiento y limpieza de datos

Los datos simulados serán sometidos a un proceso de preprocesamiento que incluirá:

- Detección y manejo de valores atípicos e inconsistentes.
- Imputación de valores faltantes según patrones observados en estudios reales.
- Codificación de variables categóricas y normalización de variables numéricas.
- Análisis exploratorio para identificar correlaciones y patrones preliminares.

Este proceso es crucial para asegurar la calidad y fiabilidad de los datos que alimentarán el modelo predictivo, aun en un contexto simulado como el presente.

3.4.5. Análisis de datos tentativo

Se realizará un análisis descriptivo y exploratorio de las variables, empleando medidas de tendencia central, dispersión y visualizaciones gráficas. Este análisis permitirá identificar tendencias hipotéticas, relaciones entre variables y posibles agrupamientos dentro de la muestra simulada. Se espera, por ejemplo, observar que estudiantes con mayor asistencia y apoyo familiar presentan mejores promedios académicos, en línea con la literatura consultada.

3.4.6. Aplicación del modelo predictivo XGBoost

Luego del análisis exploratorio, se procederá a la aplicación del modelo XGBoost, siguiendo las mejores prácticas para el entrenamiento y validación de modelos de aprendizaje automático:

- División de los datos simulados en conjuntos de entrenamiento (80%) y prueba (20%).
- Ajuste de hiperparámetros mediante validación cruzada para optimizar el rendimiento del modelo.
- Entrenamiento del modelo para predecir el rendimiento académico (clasificación y/o regresión).
- Evaluación del modelo utilizando métricas como precisión, recall, F1-score y error absoluto medio (MAE).
- Análisis de la importancia de variables para interpretar los factores más influyentes en la predicción.

3.4.7. Consideraciones éticas y validación de la veracidad de los datos

Para garantizar que los datos recolectados sean entregados de manera éticamente válida y verídica, se contemplan las siguientes estrategias:

- **Consentimiento informado claro:** Se explicará a los estudiantes y sus tutores la importancia de responder con honestidad, aclarando el anonimato y el uso exclusivo para fines académicos.

- **Garantía de anonimato:** Separación estricta de los datos personales de los datos académicos. Los datos personales (edad, género, etc.) se recolectarán y almacenarán en bases separadas de las notas y calificaciones proporcionadas por la institución educativa, evitando la vinculación directa que permita identificar individualmente a los estudiantes.

- **Validación cruzada:** Cuando sea posible, las notas y datos académicos serán obtenidos directamente del colegio, evitando que los estudiantes los reporten, lo que reduce la falsificación de información.

- **Diseño de preguntas más elaboradas:** Se incluyen preguntas con opciones escaladas y abiertas para facilitar la detección de inconsistencias o datos poco creíbles durante el análisis.

- **Aplicación supervisada:** Las encuestas serán aplicadas en condiciones controladas (en aula con presencia de facilitadores), disuadiendo la deshonestidad en las respuestas.

- **Análisis estadístico de coherencia:** Se realizará un análisis de consistencia interna y detección de patrones atípicos o incoherentes entre respuestas para identificar posibles datos falsos o poco fiables.

Estas medidas buscan proteger la integridad del estudio y la confianza en los resultados obtenidos, facilitando una interpretación más acertada del modelo predictivo desarrollado.

3.4.8. Síntesis y retroalimentación

Finalmente, los resultados obtenidos del modelo serán interpretados en función del contexto educativo cusqueño, y se propondrán recomendaciones para la gestión escolar y la política educativa local, asumiendo que la implementación práctica de este enfoque podría contribuir significativamente a la mejora del rendimiento académico en la ciudad.

Conclusiones

- La aplicación hipotética del modelo XGBoost demuestra un alto potencial para predecir el rendimiento académico en colegios secundarios de Cusco ciudad, permitiendo anticipar riesgos y orientar intervenciones educativas.
- Factores académicos, familiares y contextuales, como la asistencia, el promedio previo, el nivel educativo de los padres y el acceso a recursos tecnológicos, inciden significativamente en el desempeño estudiantil.
- La metodología propuesta es replicable y puede ser adaptada para estudios empíricos, contribuyendo a la mejora de la gestión escolar y la toma de decisiones en el ámbito educativo local.
- Se recomienda validar el modelo con datos reales y promover la digitalización y sistematización de la información educativa en Cusco, para maximizar el impacto de la inteligencia artificial en la mejora de los resultados escolares.

Recomendaciones

- Implementar sistemas de recolección digitalizada de datos académicos y socioeconómicos en colegios cusqueños.
- Capacitar a docentes y directivos en el uso de herramientas de análisis predictivo y minería de datos.
- Realizar estudios empíricos para validar y ajustar el modelo XGBoost en contextos reales.
- Promover políticas educativas que consideren factores socioeconómicos y contextuales en la gestión escolar.

Referencias Bibliográficas

- Wang, S., & Luo, B. (2024). Academic achievement prediction in higher education through interpretable modeling. *PLOS ONE, 19*(9), e0309838.
- Alba Zayas, L. E., et al. (2024). Factors influencing the academic performance of first-year medical students. *Salud, Ciencia y Tecnología – Serie de Conferencias, 3*, 662.
- Ya0̆11f0̆131, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*, 11.
- Liu, H., et al. (2022). Factors influencing secondary school students' reading literacy: An analysis based on XGBoost and SHAP methods. *Frontiers in Psychology, 13*, 948612.
- Martínez Pérez, J. R., et al. (2020). Rendimiento académico en estudiantes Vs factores que influyen en sus resultados: una relación a considerar. *EDUMECENTRO, 12*(4), 105–121.
- Hakkal, S., & Ait Lahcen, A. (2024). XGBoost to enhance learner performance prediction. *Computers and Education: Artificial Intelligence, 7*, 100254.
- Medina Canal, W. E. (2021). Hábitos de estudio y su relación con el rendimiento académico de los estudiantes, en el Área de Ciencias Sociales del Colegio Militar Pachacutec Inca Yupanqui, Cusco, 2020.
- Pfocco Huamán, S., & Pinto Valenzuela, C. (2021). Motivación y rendimiento académico en el área de matemática en estudiantes de primer grado de educación secundaria de la Institución Educativa Mixta Fortunato L. Herrera-Cusco-2020.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601-618.

- Baker, R. S. J. d., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning Analytics* (pp. 61-75). Springer, New York, NY.

- Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence, 18*(5), 411-426.

- Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*(4), 1432-1462.

- Baker, R. S. J. d. (2019). Challenges for the future of educational data mining: The Baker learning analytics prize. *Journal of Educational Data Mining, 11*(1), 1-17.

- Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review, 37*(4), 331-344.

- Delen, D., Ram, S. (2018). Predicting student performance using data mining techniques. *Journal of Educational Technology Systems, 46*(1), 3-17.

- Al-Barrak, M., & Al-Razgan, M. (2016). Predicting student grades using data mining techniques. *International Journal of Advanced Computer Science and Applications, 7*(5), 212-217.

- Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *Proceedings of 5th Annual Future Business Technology Conference*, 5-12.

A. Instrumento de Recolección de Datos

Cuestionario parte 1 (Datos sociodemográficos y Variables académicas)

Datos sociodemográficos (7 preguntas):

1. Edad: _____
2. Género: [ ] Masculino [ ] Femenino [ ] Otro [ ] Prefiero no decir
3. Grado académico: _____
4. Nivel educativo de la madre: [ ] Primaria [ ] Secundaria [ ] Superior [ ] No sabe
5. Nivel educativo del padre: [ ] Primaria [ ] Secundaria [ ] Superior [ ] No sabe
6. Tipo de gestión del colegio: [ ] Pública [ ] Privada
7. ¿Cuál es su zona de residencia? [ ] Urbano [ ] Semiurbano [ ] Rural

Variables académicas (8 preguntas):

8. Promedio de calificaciones del último año (numérico): _____
9. Porcentaje de asistencia a clases (numérico): _____ %
10. Frecuencia de estudio fuera del horario escolar: [ ] Nunca [ ] Rara vez [ ] A veces [ ] Frecuentemente [ ] Siempre
11. Participa en actividades extracurriculares: [ ] Ninguna [ ] Deportivas [ ] Culturales [ ] Académicas [ ] Otras: _____
12. Autoevaluación de hábitos de estudio: [ ] Muy malos [ ] Regulares [ ] Buenos [ ] Muy buenos
13. Tiempo promedio dedicado a estudiar diario (en minutos): _____
14. Nivel de dificultad percibida en asignaturas principales (Matemáticas, Lengua, Ciencias):
    Muy fácil [ ] Fácil [ ] Moderado [ ] Difícil [ ] Muy difícil
15. ¿Con qué frecuencia realiza tareas asignadas? [ ] Nunca [ ] Algunas veces [ ] Casi siempre [ ] Siempre

Cuestionario parte 2 (Variables familiares

Variables familiares y contextuales (6 preguntas):

16. ¿Cuenta con acceso a internet en casa? [ ] Sí, permanente [ ] Sí, ocasional [ ] No
17. ¿Tiene un lugar adecuado para estudiar en casa? [ ] Sí [ ] No [ ] A veces
18. Nivel de apoyo familiar para los estudios: [ ] Ninguno [ ] Bajo [ ] Moderado [ ] Alto
19. ¿Hay actividades distractoras frecuentes en su lugar de estudio (ruido, responsabilidades domésticas)? [ ] Sí [ ] No [ ] A veces
20. Ingreso aproximado mensual familiar: [ ] Menos de 1000 soles [ ] 1000-2000 [ ] 2001-3500 [ ] Más de 3500 [ ] Prefiero no decir
21. Número de integrantes en el hogar: _____

Percepción y motivación (4 preguntas):

22. Nivel de motivación para continuar estudios: [ ] Muy baja [ ] Baja [ ] Moderada [ ] Alta [ ] Muy alta
23. Considera importante la educación para su futuro: [ ] Nada importante [ ] Poco importante [ ] Moderadamente importante [ ] Muy importante
24. ¿Se siente apoyado por sus docentes? [ ] Nunca [ ] Algunas veces [ ] Casi siempre [ ] Siempre
25. ¿Qué tipo de dificultades académicas enfrenta? (Respuesta abierta) _____