UNIVERSIDAD CATÓLICA DE CUENCA

UNIVERSIDAD CATÓLICA DE CUENCA

Comunidad Educativa al Servicio del Pueblo

UNIDAD ACADÉMICA DE INFORMÁTICA, CIENCIAS DE LA COMPUTACIÓN E INNOVACIÓN TECNOLÓGICA

CARRERA DE TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN

ANÁLISIS DE DATOS SOCIODEMOGRÁFICOS PARA LA TOMA DE DECISIONES EN GESTIÓN EDUCATIVA DE NIVEL SUPERIOR: CASO DE ESTUDIO GRADUADOS AÑOS 2018 – 2023

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN TECNOLOGÍAS DE LA INFORMACIÓN

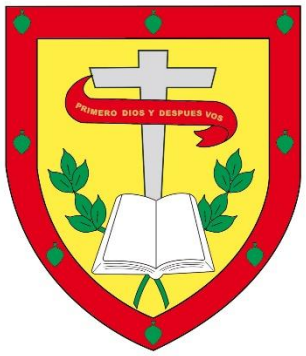
AUTOR: ERICK ANDRES BARZALLO TORRES

DIRECTOR: ING. DIANA XIMENA POMA JAPON. MSC.

CUENCA - ECUADOR

2025

DIOS, PATRIA, CULTURA Y DESARROLLO



UNIVERSIDAD CATÓLICA DE CUENCA

Comunidad Educativa al Servicio del Pueblo

UNIDAD ACADÉMICA DE INFORMÁTICA, CIENCIAS DE LA COMPUTACIÓN E INNOVACIÓN TECNOLÓGICA

CARRERA DE TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

ANÁLISIS DE DATOS SOCIODEMOGRÁFICOS PARA LA TOMA DE DECISIONES EN GESTIÓN EDUCATIVA DE NIVEL SUPERIOR: CASO DE ESTUDIO GRADUADOS AÑOS 2018 – 2023

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN TECNOLOGÍAS DE LA INFORMACIÓN

AUTOR: ERICK ANDRES BARZALLO TORRES

DIRECTOR: ING. DIANA XIMENA POMA JAPON. MSC.

CUENCA - ECUADOR

2025

DIOS, PATRIA, CULTURA Y DESARROLLO

Universidad Católica de Cuenca

DECLARATORIA DE AUTORÍA Y RESPONSABILIDAD

Declaratoria de Autoría y Responsabilidad

Erick Andrés Barzallo Torres portador(a) de la cédula de ciudadanía Nº **0106691389**. Declaro ser el autor de la obra: **"Análisis de Datos Sociodemográficos para la toma de decisiones en gestión educativa de nivel superior: Caso de estudio graduados años 2018 - 2023"**, sobre la cual me hago responsable sobre las opiniones, versiones e ideas expresadas. Declaro que la misma ha sido elaborada respetando los derechos de propiedad intelectual de terceros y eximo a la Universidad Católica de Cuenca sobre cualquier reclamación que pudiera existir al respecto. Declaro finalmente que mi obra ha sido realizada cumpliendo con todos los requisitos legales, éticos y bioéticos de investigación, que la misma no incumple con la normativa nacional e internacional en el área específica de investigación, sobre la que también me responsabilizo y eximo a la Universidad Católica de Cuenca de toda reclamación al respecto.

Cuenca, **01 de septiembre de 2025**

F: .....

Erick Andrés Barzallo Torres

C.I. **0106691389**

CERTIFICADO

Certifico que el presente trabajo titulado “ANÁLISIS DE DATOS SOCIODEMOGRÁFICOS PARA LA TOMA DE DECISIONES EN GESTIÓN EDUCATIVA DE NIVEL SUPERIOR: CASO DE ESTUDIO GRADUADOS AÑOS 2018 – 2023” fue desarrollado por Erick Andrés Barzallo Torres, bajo mi supervisión.

DIANA XIMENA POMA JAPON

Firmado digitalmente por DIANA XIMENA POMA JAPON
Fecha: 2025.09.01 10:21:32 -05'00'

F: ……..

Ing. Diana Ximena Poma Japón. Msc.

TUTOR DEL TRABAJO DE TITULACIÓN UNIVERSIDAD CATÓLICA DE CUENCA.

Dedicatoria

A mis padres y hermanos,
cimiento y refugio en cada paso,
por enseñarme con su ejemplo que la perseverancia y la unidad
son la fuerza que sostiene cualquier sueño.

A mis tíos Humberto y Zoila (+),
faros discretos en mares de duda,
que con su sencillez y cariño dejaron huellas imborrables.
Ella, desde la memoria, sigue iluminando mi camino,
recordándome que el amor permanece más allá de la ausencia.

A Nicolás, Dominga y Apolo,
fuentes inagotables de alegría,
que con sus gestos sencillos y compañía constante
llenaron de vida y ternura mis días de esfuerzo.

Y a Molly (+) y Alaska (+),
que ya no están,
pero siguen llegando primero a la puerta de mis recuerdos,
pues el lazo que formamos trasciende el tiempo
y habita en cada rincón de mi memoria.

Agradecimientos

A mis padres y hermanos,
por su amor incondicional, por acompañarme con paciencia
y por sostenerme en cada etapa de este proceso.
Sin su apoyo constante, este logro no habría sido posible;
ustedes son la base firme sobre la cual he podido construir mis sueños.

A mis jefes y al equipo de trabajo,
quienes me brindaron la oportunidad y la confianza
para seguir creciendo personal y profesionalmente,
aun cuando el tiempo parecía limitado.
Gracias por su comprensión y por enseñarme, desde la práctica,
que el esfuerzo y la disciplina siempre abren caminos.

A mis tutores, Diana y José,
por su guía cercana y paciente,
por la claridad con la que orientaron cada paso
y por motivarme a mantener siempre el compromiso académico.
Su apoyo fue fundamental para dar forma a este trabajo
y para recordarme que el conocimiento cobra sentido cuando se comparte.

A mis docentes y a la institución universitaria,
que me brindaron las herramientas necesarias
para crecer en lo académico y lo personal.
Gracias por sembrar en mí el deseo de investigar,
de aprender constantemente
y de mirar la educación como un camino de transformación.

Finalmente, a mis amigos y compañeros,
quienes con palabras de ánimo, compañía y consejos
hicieron más ligero este trayecto,
convirtiendo el esfuerzo en una experiencia compartida y significativa.

A todos ustedes,
mi gratitud sincera, profunda y eterna.

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

Análisis de datos sociodemográficos para la toma de decisiones en gestión educativa de nivel superior: Caso de estudio: graduados años 2018–2023

Sociodemographic data analysis for decision-making in higher education management: Case study: graduates from 2018–2023

Erick Barzallo-Torres[1], Diana Poma-Japón[1], José Baculima-Suárez[1]

[1]Universidad Católica de Cuenca, Cuenca, Ecuador

erick.barzallo.89@est.ucacue.edu.ec
https://orcid.org/0009-0000-1778-0021

dpomaj@ucacue.edu.ec
https://orcid.org/0000-0001-9231-1655

jbaculima@ucacue.edu.ec
https://orcid.org/0000-0002-6695-665X

Correspondencia: jbaculima@ucacue.edu.ec

Recibido: 27/04/2025
Aceptado: 29/07/2025
Publicado: 30/08/2025

Resumen

Este estudio analizó datos sociodemográficos para apoyar la toma de decisiones en gestión educativa de nivel superior en Ecuador, con énfasis en graduados de 2018–2023. El objetivo consistió en identificar los factores que incidieron en el acceso universitario y estimar su poder predictivo para la planificación de la oferta. Se utilizó un enfoque observacional y se siguió la metodología CRISP-DM. Se combinaron cuatro fuentes oficiales: graduados de bachillerato, aceptaciones en universidades, indicadores laborales y nacidos vivos, organizadas por período y provincia. Después de limpiar y estandarizar los datos, se utilizó K-Nearest Neighbors (k=3, pesos por distancia, métrica euclidiana) para modelar el acceso. Se aplicó validación cruzada (k=10) y se dividieron los datos en un 70% para entrenamiento y un 30% para prueba. Los resultados indicaron que la educación previa y el tamaño del grupo fueron los factores más importantes: el número total de estudiantes promovidos y los

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

nacimientos explicaron la mayor parte de las diferencias, mientras que los indicadores laborales no tuvieron ninguna importancia en el modelo. El desempeño alcanzó R^2=0.701, MAE=1 376.50 y RMSE=2 247.21, con sesgo de subestimación en valores extremos y heterocedasticidad en altas demandas. Por lo tanto, las conclusiones sugirieron que la planificación debería enfocarse en la retención y promoción en el bachillerato, utilizando proyecciones demográfico-educativas para determinar la cantidad de cupos. Además, es importante no olvidar que hay variables institucionales y familiares que no se ven y que pueden afectar los picos de acceso. Se sugirió, para futuras investigaciones, incluir datos individuales, indicadores de calidad escolar y comparaciones entre provincias, así como utilizar enfoques mixtos que evalúen el impacto cultural y los resultados académicos, para mejorar la toma de decisiones basadas en evidencia.

Palabras clave: Educación superior, análisis de datos, demografía, indicadores educativos, toma de decisiones.

Abstract

This study analyzed sociodemographic data to support decision-making in higher education management in Ecuador, with an emphasis on graduates from 2018–2023. The objective was to identify the factors that influenced university access and to estimate their predictive power for supply planning. An observational approach was used, and the CRISP-DM methodology was followed. Four official sources were combined: high school graduates, university admissions, labor market indicators, and live births, organized by period and province. After cleaning and standardizing the data, K-Nearest Neighbors (k=3, distance-weighted, Euclidean metric) was used to model access. Ten-fold cross-validation was applied, and the data were split into 70% for training and 30% for testing. The results indicated that prior education and group size were the most important factors: the total number of students promoted and births accounted for most of the variance, while labor market indicators had no significance in the model. The performance reached R^2 = 0.701, MAE = 1,376.50, and RMSE = 2,247.21, with underestimation bias at extreme values and heteroscedasticity at high demand levels. Therefore, the conclusions suggested that planning should focus on retention and promotion in upper secondary education, using demographic-educational projections to determine the number of slots. Furthermore, it is important not to forget that

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

there are unseen institutional and family variables that can affect access peaks. It was suggested that future research include individual data, school quality indicators, and comparisons between provinces, as well as employ mixed-methods approaches that assess cultural impact and academic outcomes, to improve evidence-based decision-making.

Keywords: Higher education, data analysis, demography, educational indicators, decision-making.

Introducción

En Ecuador, el acceso a la educación superior continúa siendo un desafío para miles de estudiantes que culminan el bachillerato, especialmente en contextos rurales y de bajos recursos. Según [1], esta problemática persiste debido a la desigualdad estructural que limita las oportunidades en áreas marginales. Por su parte, [2] destacan que el entorno socioeconómico, como también la falta de políticas públicas efectivas, influyen negativamente en la continuidad académica. De igual manera, [3] argumentan que las condiciones familiares, como también el desempleo juvenil, dificultan el ingreso a la educación superior, incrementando así brechas históricas en el sistema educativo.

Estudios recientes del GEM Report 2023 de la UNESCO [4] revelan que el 52 % de las diferencias en culminación de secundaria a nivel mundial están influenciadas por condiciones socioeconómicas y demográficas. Por su parte, la [5] advierte que en América Latina un incremento del 1% en el desempleo juvenil reduce la matrícula universitaria en 0,4%, siendo más crítico en provincias rurales. Estas evidencias destacan la importancia de incluir variables laborales y demográficas en los modelos que analizan el ingreso a la universidad.

Sin embargo, en el contexto ecuatoriano aún no se dispone de un sistema integrado que combine información académica, económica y sociodemográfica para explicar las dinámicas de acceso universitario con precisión territorial y temporal. Esto genera una brecha informativa que limita la planificación de políticas públicas más equitativas.

Este estudio tiene como objetivo identificar los factores que inciden en el acceso a la educación superior en Ecuador entre 2018 y 2023. Para ello, se integraron cuatro fuentes

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

oficiales de datos sobre nacimientos, empleo, estudiantes promovidos de educación secundaria y aceptación universitaria, utilizando la metodología CRISP-DM y el modelo K-Nearest Neighbors para predecir el comportamiento del acceso a la educación superior.

El enfoque propuesto permite revelar relaciones consistentes entre las variables sociodemográficas y el acceso a la educación superior. Este análisis aporta información relevante para la planificación educativa y puede servir como base para el desarrollo de estrategias de mejora del tránsito entre la educación media y la superior.

Diversas metodologías han sido desarrolladas para orientar proyectos de minería de datos, especialmente en contextos educativos. Entre las más reconocidas destacan CRISP-DM, SEMMA (Sample, Explore, Modify, Model, Assess) y KDD (Knowledge Discovery in Databases), cada una con enfoques particulares que influyen en la planificación, ejecución e interpretación del análisis de datos.

CRISP-DM es la metodología más ampliamente adoptada en investigaciones con datos educativos debido a su estructura iterativa de seis fases, que conecta los objetivos institucionales con tareas técnicas específicas. Se valora por su trazabilidad, claridad documental y compatibilidad con herramientas de código abierto como Python y R. Más del 70% de los estudios recientes la utilizan por su robustez en entornos con grandes volúmenes de datos [6], [7], [8], [9], [10].

SEMMA, desarrollada por el SAS Institute, se centra en la ingeniería de variables y el modelado analítico. Su flujo técnico de cinco etapas facilita la transformación de macrodatos y ha sido útil en aplicaciones de aprendizaje profundo. Sin embargo, su dependencia del ecosistema SAS limita su accesibilidad en contextos educativos con restricciones presupuestarias [11], [12], [13].

KDD, en cambio, funciona como un marco conceptual integral que abarca desde la selección y transformación de datos hasta la extracción e interpretación de patrones. Su flexibilidad lo convierte en una base teórica sólida para proyectos basados en big data y análisis distribuido, aunque requiere una adecuada especificación de fases para evitar ambigüedades en su implementación [14], [15], [16].

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

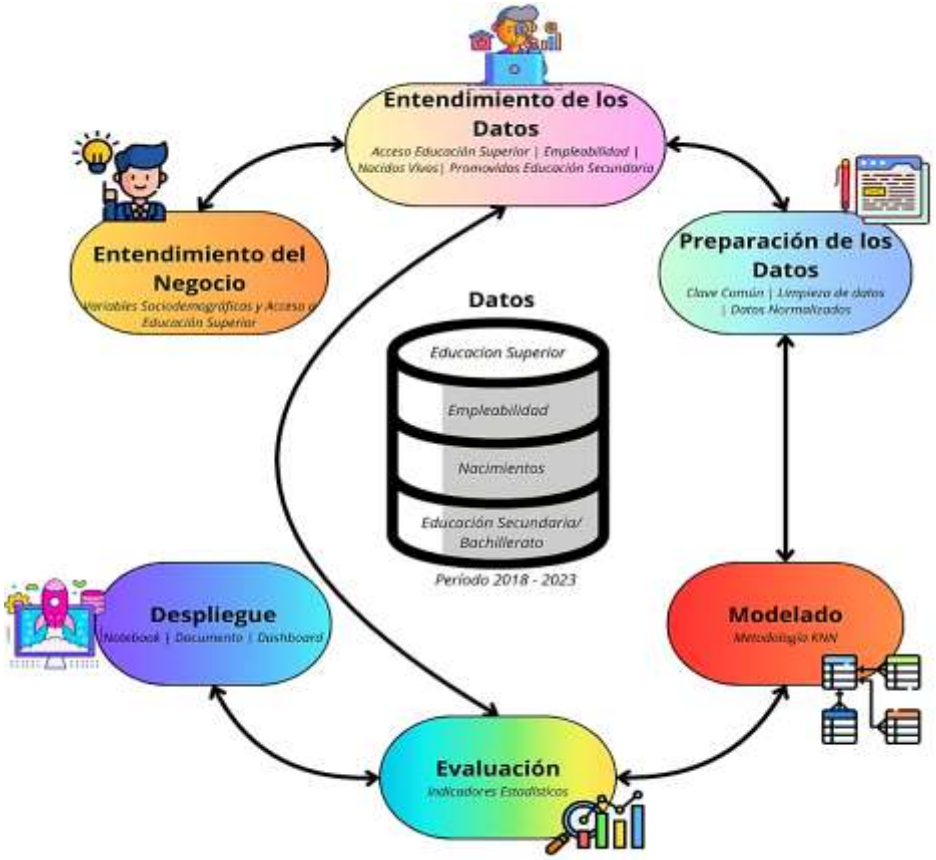
En función de estas características, la presente investigación adopta CRISP-DM como metodología principal por su equilibrio entre estructura formal y adaptabilidad técnica, lo que la convierte en una opción idónea para proyectos educativos basados en datos abiertos y necesidades institucionales concretas.

Metodología

Se realizó un estudio observacional retrospectivo con enfoque predictivo, utilizando la metodología CRISP-DM para minería de datos educativos, por su capacidad de alinear los objetivos institucionales con decisiones técnicas de análisis. Esta metodología organiza el proceso en seis fases iterativas que garantizan una ejecución ordenada, trazable y adaptable a los proyectos educativos. La Figura 1 ilustra este ciclo aplicado al estudio.

Figura 1

Fases iterativas de CRISP-DM



ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

En la fase de entendimiento del negocio se definen los objetivos estratégicos, la disponibilidad de recursos y el plan de proyecto, traducidos en metas de minería de datos [9]. En la fase de entendimiento de los datos se recopilan las fuentes, se describe su estructura, se exploran relaciones y se verifica la calidad antes de cualquier transformación [10]. En la preparación de los datos se selecciona atributos, se realiza la limpieza de valores anómalos, se integran las diferentes bases de datos, todas las variables predictoras se escalan mediante StandardScaler y se formatea el conjunto final "datos integrados" [9]. Para el modelado se usa el algoritmo KNN Regressor, con optimización de hiperparámetros mediante grid search: k=3 vecinos (evaluado rango 3-21 con incrementos de 2), pesos por distancia, métrica euclidiana. Se realiza la evaluación del modelo empleando validación cruzada con k-fold (k=10) y dividiendo los datos: 70% entrenamiento y 30% prueba; empleando métricas de desempeño R^2 (coeficiente de determinación), que mide la varianza explicada por el modelo; MAE (Error Absoluto Medio), que representa el error promedio en unidades de la variable objetivo; RMSE (Raíz del Error Cuadrático Medio), que cuantifica el error con mayor penalización para valores atípicos; y MAPE (Error Porcentual Absoluto Medio), que expresa el error relativo en términos porcentuales. La importancia de variables se determinó por permutation importance (100 permutaciones, IC 95%). Se contrasta el modelo con los criterios de negocio, se validan supuestos y se decide si los hallazgos son suficientemente sólidos para producción [10]. Finalmente, en el despliegue se implementa el modelo, se planifica su mantenimiento, se redacta el informe final y se realiza la retrospectiva del proyecto para mejoras futuras [10].

Las variables y datos se obtuvieron de 4 fuentes:

- Promovidos Educación Secundaria (MinEduc, Ecuador) períodos 2018-2023
- Aceptaciones universitarias (SENESCYT, Ecuador), período 2018-2023
- Empleabilidad, Indicadores laborales (INEC, Ecuador), período 2018-2023
- Registro demográfico de nacimientos (INEC, Ecuador), período 2000-2005

Los datos fueron agrupados por período y por provincia, a excepción de los datos de empleabilidad, que fueron agrupados solamente por período, al no encontrarse la variable provincia en la base de datos.

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

La variable dependiente fue Accesos a Educación Superior (estudiantes aceptados en la universidad). Las variables predictoras fueron: 'Total Estudiantes Promovidos', 'Estudiantes Femenino Promovidos Tercer Año Bachillerato', 'Estudiantes Masculino Promovidos Tercer Año Bachillerato', 'Año Nacimiento', 'Nacimientos', 'Tasa de desempleo (%', 'Tasa de empleo adecuado (%', 'Tasa de subempleo (%', 'Tasa de empleo (%', 'Sector informal (%', 'Promedio de horas trabajadas', 'Promedio Ingreso Laboral ($)'.

Para el análisis y desarrollo de la metodología se usó el entorno de desarrollo Jupyter Lab y el lenguaje de programación Python con librerías estándar. Para el algoritmo KNN se usó la librería scikit-learn.

Resultados

Los resultados de este estudio se presentan a continuación de acuerdo a las fases ejecutadas de la metodología CRISP-DM.

Entendimiento del negocio:

1. Objetivo del negocio: Identificar los factores sociodemográficos y académicos que influyen en el acceso a la educación superior. El fin es mejorar la planificación educativa y modelar el acceso a la educación superior en Ecuador. Además, se busca identificar los factores más influyentes en la transición del bachillerato a la universidad.
2. Situación actual. Se requiere comprender las tendencias en la educación superior y su relación con variables sociodemográficas para mejorar la planificación de la oferta académica y las políticas de acceso a la educación. Actualmente, este análisis se realiza de manera fragmentada y con reportes estáticos que no permiten una visión integral ni predictiva.

   Para sustentar este análisis, se han identificado y obtenido cuatro fuentes de datos abiertas: aceptaciones universitarias por provincia y semestre del Sistema de Información Académica Unificada de la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SIAU - SENESCYT) [17], indicadores semestrales de empleo publicados en la Encuesta Nacional de Empleo, Desempleo y Subempleo del

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

Instituto Nacional de Estadística y Censos (INEC) [18], registro de nacidos vivos por provincia del Instituto Nacional de Estadística y Censos (INEC) [19] y la base de graduados de bachillerato (promovidos de educación secundaria) del Ministerio de Educación [20].

3. Objetivos de minería de datos: fusionar las bases por período y provincia y construir un modelo que pronostique el acceso a la educación superior.
4. Plan del proyecto. El proyecto se desarrollará siguiendo las fases de la metodología CRISP-DM, donde se fusionarán las diferentes bases de datos y se modelará con K-Nearest Neighbors Regression usando el lenguaje Python; los resultados permitirán obtener una visión del acceso a la educación superior de acuerdo a variables sociodemográficas.

Entendimiento de los datos:

1. Recolección de datos iniciales. Se consolidaron cuatro conjuntos públicos, cada uno con la clave período y provincia, que incluyen las variables necesarias para el análisis posterior. Estos datos se presentan en la Tabla 1, donde se detallan las fuentes, dimensiones y características principales de cada conjunto.

Tabla 1

Datos Iniciales y Variables

Conjunto	Variables retenidas	Descripción breve
Aceptaciones universitarias [17]	Período, semestre, provincia	Cupos aceptados por provincia en cada semestre (2018-2023).
Indicadores de empleo [18]	Período, semestre, Tasa de desempleo (%), Tasa de empleo adecuado (%), Tasa de subempleo (%), Tasa de empleo (%), Sector informal (%), Promedio de horas trabajadas, Promedio Ingreso Laboral ($)	Series semestrales 2018-2023; tasas promediadas y métricas convertidas a decimales para relacionar mercado laboral con ingreso universitario.
Nacidos vivos [19]	Año, provincia	Registros anuales 2000-2004; finalizan educación secundaria (bachillerato) 18 años después de 2018 a 2023.
Promovidos educación secundaria (Graduados de bachillerato) [20]	Periodo, Semestre, Provincia, Estudiantes Femenino Promovidos Tercer Año Bachillerato, Estudiantes Masculino Promovidos Tercer Año Bachillerato	19581 registros 2018-2023.

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

Preparación de los datos:

1. Selección de variables claves. Se retuvieron 15 campos: Periodo, Provincia, Total Estudiantes Promovidos, Estudiantes Femenino Promovidos Tercer Año Bachillerato, Estudiantes Masculino Promovidos Tercer Año Bachillerato, Accesos Educación Superior, Año Nacimiento, Nacimientos, Tasa de desempleo (%, Tasa de empleo adecuado (%, Tasa de subempleo (%, Tasa de empleo (%, Sector informal (%, Promedio de horas trabajadas, Promedio ingreso laboral ($). Estos campos se mantienen luego de integrar los cuatro conjuntos de datos y cubrir las dimensiones académicas, empleabilidad y demográfica (véase Tabla 2).

2. Transformaciones

Tabla 2

Variables comunes y pasos de depuración

Conjunto	Pasos de depuración y estandarización
Promovidos Educación Superior (Graduados de bachillerato)	Eliminar 231 columnas administrativas Descartar filas con conteos vacíos, nulos o cero
Aceptaciones universitarias	Borrar filas totalmente vacías Mantener las 24 columnas de provincia
Indicadores de empleo	Agregar registros semestrales Convertir/Estandarizar tasas, horas (hh:mm) e ingreso (USD) Etiquetar unidades y suprimir filas sin datos numéricos.
Nacidos vivos	Eliminar filas con todos los campos vacíos/nulos Acotar a 2000-2004 para corresponder a la cohorte que es promovida de educación secundaria y accede a la educación superior en 2018-2023.

3. Resultado. Las cuatro tablas limpias comparten la clave período y tres tablas (Promovidos Educación Superior, Aceptaciones Universitarias, Nacimientos) comparten las claves período y provincia; este dataset integrado alimentará las fases de modelado y evaluación.

Modelado. Para identificar el modelo de k-Vecinos Más Cercanos (KNN) con el mejor desempeño predictivo, se realizó una búsqueda exhaustiva de hiperparámetros mediante validación cruzada. El proceso de optimización se centró en el número de vecinos (n_neighbors), la función de peso (weights) y la métrica de distancia (metric). El análisis determinó que la configuración óptima que maximizó el coeficiente de determinación (R^2) fue la siguiente: Número de vecinos (k) = 3, Función de peso = 'distance' (los vecinos más

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

cercanos tienen una mayor influencia en la predicción), Métrica de distancia = Distancia euclidiana ('euclidean'). Con esta combinación de parámetros, el modelo alcanzó un score de validación cruzada (R^2) de 0.762, lo que indica un buen poder predictivo y una robustez considerable en los subconjuntos de entrenamiento y validación. Posteriormente, se entrenó un modelo final utilizando todos los datos de entrenamiento y la configuración óptima identificada. Al evaluar este modelo final sobre el conjunto de entrenamiento, se obtuvo un coeficiente de determinación (R^2) de 0.701. Este valor confirma la capacidad del modelo para explicar una porción significativa (70.1%) de la varianza de la variable dependiente en los datos de entrenamiento.

La ligera discrepancia entre el score de validación cruzada (0.762) y el score en el conjunto de entrenamiento (0.701) es esperable. El score de validación cruzada representa un promedio del rendimiento en múltiples particiones de validación, mientras que el score reportado en el modelo final es una medida puntual sobre el conjunto completo de entrenamiento. La concordancia entre ambas métricas sugiere que el modelo está bien ajustado y no presenta un sobreajuste severo.

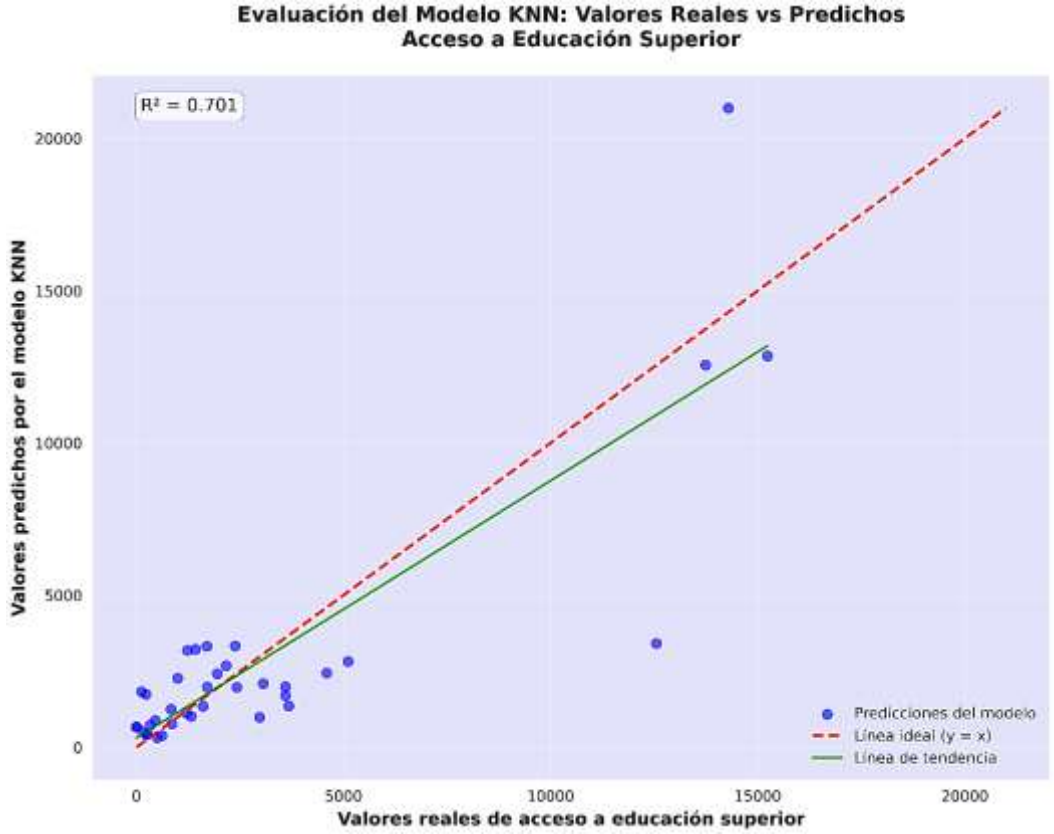
Las variables predictoras utilizadas fueron: Periodo, Provincia, Total Estudiantes Promovidos, Estudiantes Femenino Promovidos Tercer Año Bachillerato, Estudiantes Masculino Promovidos Tercer Año Bachillerato, Año Nacimiento, Nacimientos, Tasa de desempleo (%, Tasa de empleo adecuado (%, Tasa de subempleo (%, Tasa de empleo (%, Sector informal (%, Promedio de horas trabajadas y Promedio Ingreso Laboral ($).

Evaluación. La Figura 2 muestra que existe una relación clara entre los valores reales y los predichos por el modelo, aunque con algunas variaciones en casos extremos. El modelo KNN mostró un ajuste bueno (coeficiente de determinación $R^2 = 0.701$), explicando aproximadamente el 70% de la variabilidad en el acceso a educación superior. Sin embargo, se observa un sesgo sistemático de subestimación, evidenciado por la desviación consistente de la línea de tendencia por debajo de la línea ideal (y = x). La dispersión de los residuos sugiere heterocedasticidad, con mayor variabilidad en las predicciones para valores superiores a 10000.

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

Figura 2

Dispersión entre valores reales y predichos



La Figura 3 muestra el análisis de predicciones del modelo KNN Regressor, comparando los valores reales versus los predichos de acceso a educación superior para el conjunto de prueba (35 observaciones). El gráfico revela el comportamiento predictivo del modelo en observaciones individuales, donde las áreas sombreadas indican las direcciones de los errores de predicción.

Los resultados evidencian un patrón mixto de precisión predictiva, demostrando capacidad para capturar la tendencia general de los datos, particularmente en observaciones con valores bajos a moderados (índices 1-8 y 30-35). Sin embargo, se identificaron limitaciones significativas en la predicción de valores extremos, especialmente en las observaciones 10, 16, 24 y 27, donde se registraron discrepancias considerables entre valores reales y predichos.

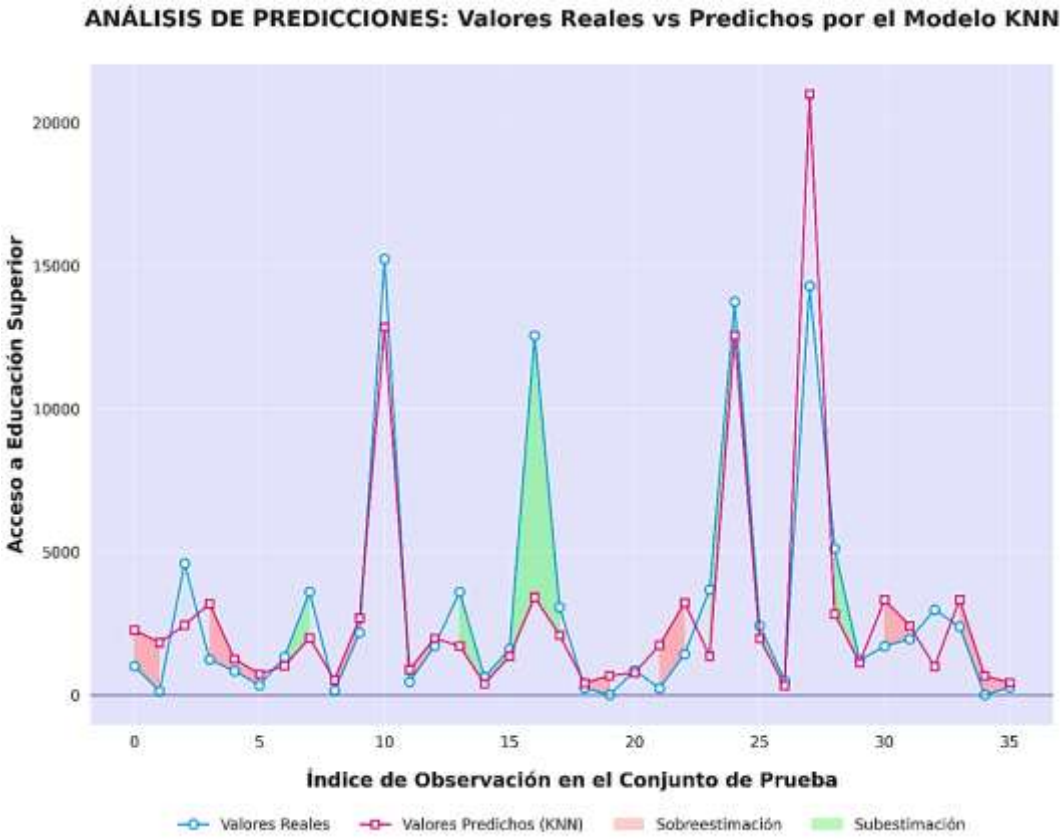
El análisis visual confirma la presencia de tanto subestimación (áreas verdes) como sobreestimación (áreas rosadas) en las predicciones. El modelo mostró dificultades para

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

predecir con precisión los picos de acceso más altos, como se ve en las observaciones 10 y 27, sugiriendo limitaciones en la captura de patrones en los extremos superiores de la distribución. Estos hallazgos complementan los resultados del coeficiente de determinación ($R^2 = 0.701$).

Figura 3

Comparación de valores reales y predichos (primeros 36 casos)



La Figura 4 presenta el análisis de importancia de variables del modelo KNN mediante el método de permutación con 100 iteraciones e intervalos de confianza del 95%. Los valores representan la disminución en R^2 cuando cada variable es permutada aleatoriamente, indicando su contribución relativa al poder predictivo del modelo.

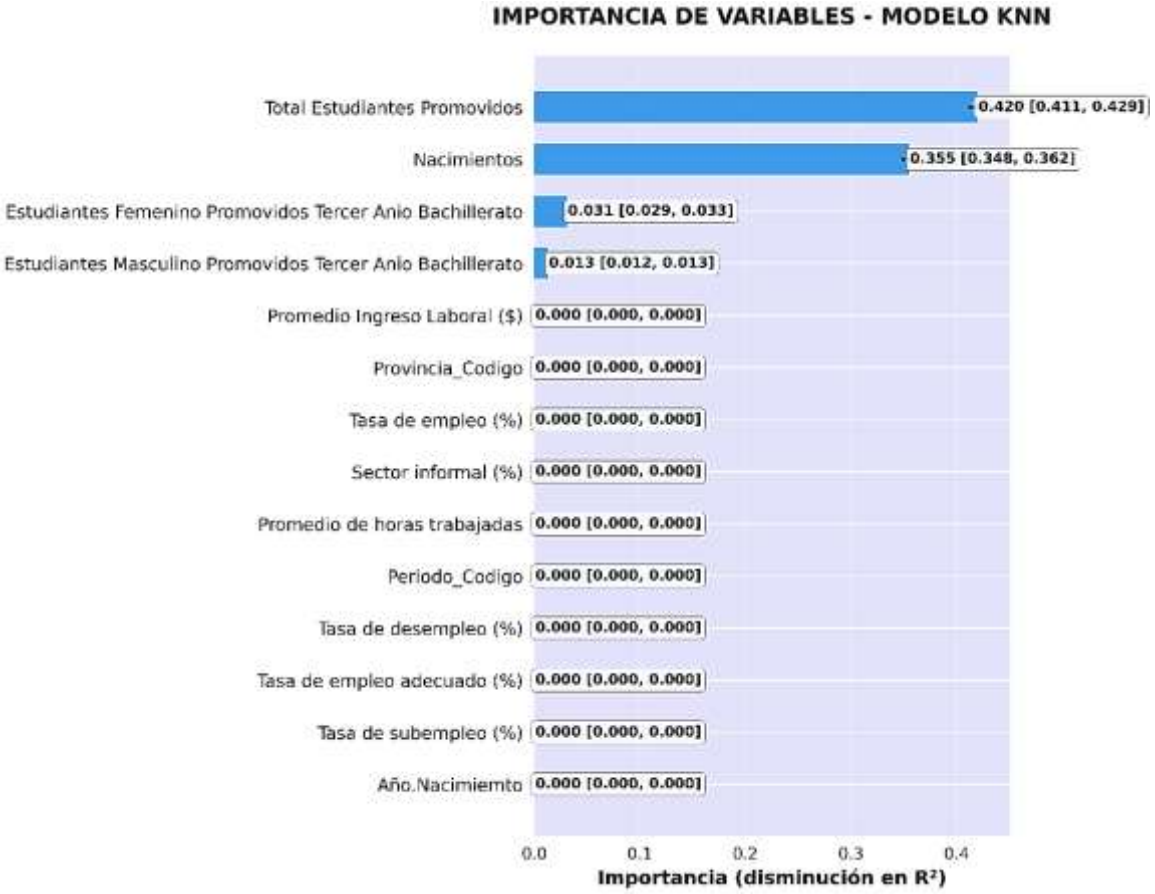
Los resultados muestran que cuatro variables son importantes para predecir el acceso a la educación superior. Las variables "Estudiantes promovidos a educación secundaria" y "Nacimientos" juntas explican la mayor parte del poder predictivo del modelo. Las variables socioeconómicas incluidas en el modelo mostraron importancia nula (0.000), indicando que

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

no contribuyen al poder predictivo del modelo KNN bajo las condiciones específicas de este análisis.

Figura 4

Importancia de variables determinada por permutation importance (100 permutaciones, IC 95%)

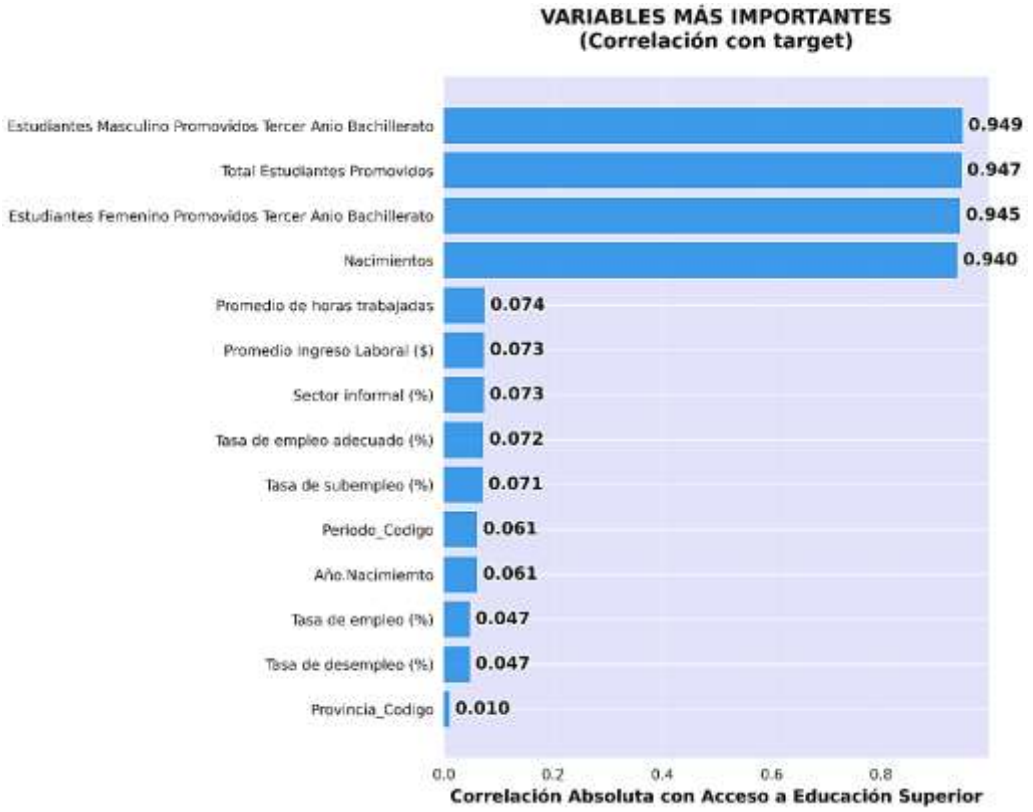


La Figura 5 presenta el análisis de la relación directa entre las variables que predicen y la variable objetivo (acceso a educación superior), organizadas según la fuerza de su asociación. Este análisis complementa los resultados de importancia de variables del modelo KNN. Los resultados revelan un patrón de correlaciones fuertemente polarizado. Las cuatro variables educativo-demográficas principales presentaron correlaciones altas con el acceso a educación superior.

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

Figura 5

Correlación entre las variables predictoras y la variable acceso a educación superior



Finalmente, el desempeño del modelo K-Nearest Neighbors Regressor con k = 3 alcanzó un R^2 de 0.701, MAE de 1376.50 y RMSE de 2247.21 (ver Tabla 3).

Tabla 3

Desempeño del modelo KNN Regressor

Métricas	Valor
Error absoluto medio (MAE)	1376.50 cupos
Raíz del error cuadrático medio (RMSE)	2247.21 cupos
Coeficiente de determinación (R^2)	0.701

Despliegue. Los resultados obtenidos permiten generar un informe orientado a la toma de decisiones educativas, especialmente en relación con el acceso a la universidad. Esta información puede ser utilizada por instituciones y autoridades para diseñar estrategias que garanticen una transición más equitativa entre la educación secundaria y la universidad.

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

Discusión

Los resultados del modelo KNN ($R^2 = 0.701$) indican patrones claros en los determinantes del acceso a educación superior, con implicaciones importantes para la comprensión de las transiciones educativas, y sugieren que aproximadamente el 70% de la variabilidad en el acceso universitario puede explicarse mediante las variables incluidas, lo cual representa un nivel de precisión aceptable para fenómenos sociales complejos, donde las variables educativo-demográficas son predictores dominantes. Las correlaciones superiores a 0.94 entre el acceso universitario y los estudiantes promovidos de bachillerato, junto con la alta importancia de "Total Estudiantes Promovidos" y "Nacimientos", confirman que el flujo educativo previo y el tamaño de cohorte son los determinantes primarios.

Los resultados demuestran que el uso de técnicas de minería de datos y el algoritmo KNN Regressor permiten modelar de forma eficaz el acceso a la educación superior a partir de variables como número de promovidos y nacimientos. Esto es consistente con lo señalado por [21], quienes aplicaron minería de datos a bases educativas ecuatorianas para predecir el rendimiento académico en línea, resaltando el valor de integrar información sociodemográfica y académica. Asimismo, estudios como el de [22] han señalado que la aplicación de técnicas de minería educativa no solo facilita la predicción, sino que también mejora los procesos de planificación institucional.

Por otro lado, la nula importancia de las variables socioeconómicas territoriales (ingresos, empleo, desempleo) contrasta con expectativas teóricas basadas en teorías de capital humano. Esta ausencia sugiere dos interpretaciones: que las condiciones socioeconómicas agregadas no reflejan las decisiones individuales familiares, o que políticas de democratización educativa han reducido efectivamente las barreras económicas al acceso universitario.

El sesgo sistemático de subestimación y la dificultad para predecir valores extremos indican factores no capturados que impulsan el acceso a la educación superior por encima de las predicciones demográfico-educativas. La heterocedasticidad observada sugiere que los mecanismos de acceso operan diferentemente según el contexto de demanda.

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

Los resultados sugieren que las políticas más efectivas para incrementar el acceso universitario deberían focalizarse en la retención y promoción en bachillerato, más que en intervenciones socioeconómicas generales. La predominancia de variables educativas directas implica que el acceso a educación superior depende fundamentalmente de la calidad y continuidad del pipeline educativo previo.

Es necesario que las unidades responsables de la planificación educativa fortalezcan sus competencias en análisis de datos, tal como lo recomiendan [23] en su revisión sobre minería educativa. El uso adecuado de estas metodologías permitirá no solo detectar patrones, sino también generar estrategias focalizadas para reducir las brechas en el acceso universitario, con base en evidencia y segmentación precisa.

Conclusiones

Este estudio demostró la efectividad del algoritmo K-Nearest Neighbors Regressor para modelar el acceso a educación superior en Ecuador utilizando datos integrados de fuentes oficiales. El modelo final (k=3) alcanzó un desempeño satisfactorio (R^2 = 0.701, MAE = 1376.50, RMSE = 2247.21), explicando el 70% de la variabilidad en el acceso universitario provincial.

Los resultados confirman que las variables educativo-demográficas son los predictoras dominantes. Los estudiantes que pasaron de bachillerato (educación secundaria) y el tamaño de la población de su año (nacimientos) tuvieron correlaciones de más de 0.94 y fueron estadísticamente importantes, mientras que las variables socioeconómicas de la zona no mostraron ninguna contribución. Esto indica que el pipeline educativo previo es más determinante que las condiciones económicas agregadas para explicar el acceso universitario.

Los hallazgos orientan hacia estrategias centradas en la retención y calidad del bachillerato como mecanismo más efectivo para incrementar el acceso universitario. La metodología desarrollada permite anticipar la demanda de cupos basándose en indicadores demográficos-educativos, facilitando una planificación más precisa de la oferta académica.

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

El sesgo de subestimación sistemática indica factores no capturados en el modelo. Investigaciones futuras requieren datos individuales, variables de calidad educativa institucional y factores culturales familiares para capturar completamente los mecanismos de decisión sobre educación superior. Además, deben aplicar modelos de clasificación supervisada que permitan predecir no solo el acceso, sino también la permanencia y el éxito académico en la educación superior.

Este trabajo establece la aplicación de minería de datos en decisiones educativas basadas en evidencia, demostrando el valor de integrar múltiples fuentes oficiales mediante metodologías CRISP-DM para generar conocimiento sobre transiciones educativas en Ecuador.

Referencias

[1] E. A. Gallardo-Lara (Universidad Técnica de Babahoyo, UTB), "Análisis general del acceso al sistema de educación superior del Ecuador 2009–2018," *Magazine de las Ciencias: Revista de Investigación e Innovación*, vol. 8, no. 1, pp. 21–37, Jan. 2023, doi: 10.33262/RMC.V8I1.974.

[2] E. Stefos (Universidad Nacional de Educación, UNAE) and C. E. Chávez Morales (Universidad de Las Américas, UDLA), "Brechas educativas en Ecuador: el caso de la población con estudios universitarios," *Revista Scientific*, vol. 8, no. 28, pp. 230–244, May–Jul. 2023, doi: 10.29394/Scientific.issn.2542-2987.2023.8.28.12.230-244.

[3] D. A. Morales Echeverria and S. R. Galarza Arrieta, "Acceso a la educación superior pública en Ecuador y limitación de cupos como vulneración de derechos," *Revista de Investigación Educativa Niveles*, vol. 1, no. 2, pp. 5–13, Jul. 2024, doi: 10.61347/rien.v1i2.60.

[4] UNESCO (United Nations Educational, Scientific and Cultural Organization), *Global Education Monitoring Report 2023: Technology in Education: A Tool on Whose Terms?*, Paris: UNESCO, 2023. [Online]. Available: https://digitallibrary.un.org/record/4020460.

[5] Naciones Unidas (UNESCO Regional Office for Latin America and the Caribbean), *La encrucijada de la educación en América Latina y el Caribe: Informe regional de*

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

monitoreo ODS4-Educación 2030, Santiago de Chile: UNESCO, 2022. [Online]. Available: www.unesco.org/open-access/terms-use-ccbysa-sp.

[6] A. M. Shimaoka, R. C. Ferreira, and A. Goldman, "The evolution of CRISP-DM for data science: Methods, processes and frameworks," *Revista Latinoamericana de Ingeniería de Software*, 2023. [Online]. Available: https://n9.cl/nq5rld.

[7] A. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[8] M. Elkabalawy, A. Al-Sakkaf, E. M. Abdelkader, and G. Alfalah, "CRISP-DM-based data-driven approach for building energy prediction utilizing indoor and environmental factors," *Sustainability*, vol. 16, no. 17, p. 7249, Sep. 2024, doi: 10.3390/su16177249.

[9] Data Science PM (Data Science Project Management Organization), "What is CRISP-DM?," 2023. [Online]. Available: https://www.datascience-pm.com/crisp-dm-2/.

[10] NCR and J. Clinton (DaimlerChrysler), *CRISP-DM 1.0 Step-by-Step Data Mining Guide*, 1999. [Online]. Available: https://n9.cl/rlry9h. [Accessed: Aug. 20, 2025].

[11] O. Firas, "A combination of SEMMA and CRISP-DM models for effectively handling big data using formal concept analysis based knowledge discovery: A data mining approach," *World Journal of Advanced Engineering and Technology Sciences*, vol. 8, no. 1, pp. 147–158, Jan. 2023, doi: 10.30574/WJAETS.2023.8.1.0147.

[12] Data Science PM, "What is SEMMA?," 2023. [Online]. Available: https://www.datascience-pm.com/semma/.

[13] A. Alika, H. Mirza, Andri, and Ferdiansyah, "Classification of South Sumatra songket woven fabric motifs using deep learning," *Int. J. Data Sci. Comput.*, vol. 2, no. 3, pp. 45–52, Jul. 2024, doi: 10.61978/data.v2i3.313.

[14] V. Plotnikova, M. Dumas, and F. Milani, "Adaptations of data mining methodologies: A systematic literature review," *PeerJ Computer Science*, vol. 6, no. e267, pp. 1–32, 2020, doi: 10.7717/peerj-cs.267.

[15] N. Singhal and Himanshu, "A review on knowledge discovery from databases," in *Advances in Intelligent Systems and Computing*, Singapore: Springer, 2022, pp. 563–574, doi: 10.1007/978-981-16-9488-2_43.

ISSN: 3028-8800
Julio-diciembre, 2025
Vol. 4 - No. 2
DOI: https://doi.org/10.62943/rig.v4n2.2025.342

[16] A. Kumar, S. Limon, Y. Li, S. Bold, and S. Urschel, "A knowledge discovery process extended to experimental data for the identification of motor misalignment patterns," *Machines*, vol. 11, no. 8, p. 827, Aug. 2023, doi: 10.3390/machines11080827.

[17] SENESCYT (Secretaría de Educación Superior, Ciencia, Tecnología e Innovación, Ecuador), "Acceso a la Educación Superior – Oferta y Aceptaciones – Servicios Senescyt," 2023. [Online]. Available: https://siau.senescyt.gob.ec/acceso-educacion-superior/.

[18] INEC (Instituto Nacional de Estadística y Censos, Ecuador), "Microsoft Power BI," 2023. [Online]. Available: https://n9.cl/3eqkx. [Accessed: Aug. 20, 2025].

[19] INEC (Instituto Nacional de Estadística y Censos, Ecuador), "Nacidos vivos y defunciones fetales," 2023. [Online]. Available: https://www.ecuadorencifras.gob.ec/nacidos-vivos-y-defunciones-fetales/.

[20] Ministerio de Educación del Ecuador, "Base de datos – Ministerio de Educación," 2023. [Online]. Available: https://educacion.gob.ec/base-de-datos/.

[21] J. Rodas-Silva and J. Parraga-Alava, "Predicting academic performance of low-income students in public Ecuadorian online universities: An educational data mining approach," in *Proc. 15th Int. Conf. EduTech*, INSTICC, Jul. 2023, pp. 233–240, doi: 10.5220/0012086300003541.

[22] H. Almaghrabi, B. Soh, A. Li, and I. Alsolbi, "SoK: The impact of educational data mining on organisational administration," *Information*, vol. 15, no. 11, p. 738, Nov. 2024, doi: 10.3390/info15110738.

[23] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *arXiv preprint* arXiv:2402.07956, 2024. [Online]. Available: https://arxiv.org/pdf/2402.07956.

Los autores no tienen conflicto de interés que declarar. La investigación fue financiada por la Universidad Católica de Cuenca y los autores.