A Scalable Data Mining Framework for Student Dropout Prediction in Higher Education

Rony David Trespalacios Arismendi[1]; Sara Cristina Guerrero[2]; Oscar Agudelo Varela[3]
[1,2,3]Universidad de los Llanos, Colombia, *rony.trespalacios@unillanos.edu.co, sguerrero@unillanos.edu.co, oscar.agudelo@unillanos.edu.co*

***Abstract*– This paper outlines the application of the CRISP-DM methodology to develop a data mining approach for both predicting dropout and understanding its underlying causes. Using data from the university's University Academic Information System (SIAU) and additional data sources, the project is currently in the Data Preparation phase, where an ETL (Extraction, Transformation, and Loading) process is being implemented to prepare the data for later modeling. The research aims to explore machine learning algorithms, such as decision trees and neural networks, to identify predictive patterns and reveal correlations related to dropout factors. This Work in Progress paper details the adapted CRISP-DM methodology and the planned stages of model development, evaluation, and software prototyping, with the ultimate goal to inform effective interventions and promote student success at FCBI-Unillanos.**

***Keywords*-- Educational Data Mining (EDM), CRISP-DM, Student Dropout Prediction, ETL.**

Marco de Minería de Datos Escalable para la Predicción de la Deserción Estudiantil en la Educación Superior

Rony David Trespalacios Arismendi[1]; Sara Cristina Guerrero[2]; Oscar Agudelo Varela[3]
[1,2,3]Universidad de los Llanos, Colombia, *rony.trespalacios@unillanos.edu.co, sguerrero@unillanos.edu.co, oscar.agudelo@unillanos.edu.co*

Resumen– ***Este artículo describe la aplicación de la metodología CRISP-DM para desarrollar un enfoque de minería de datos destinado tanto a predecir la deserción estudiantil como a comprender sus causas subyacentes. Utilizando datos del Sistema de Información Académico Universitario (SIAU) de la universidad y fuentes de datos adicionales. El proyecto se encuentra actualmente en la fase de Preparación de Datos, donde se está implementando un proceso ETL (Extracción, Transformación y Carga) para preparar la información para el modelado posterior. La investigación tiene como objetivo explorar algoritmos de aprendizaje automático, tales como árboles de decisión y redes neuronales, para identificar patrones predictivos y revelar correlaciones relacionadas con los factores de deserción. Este artículo de Trabajo en Progreso detalla la metodología CRISP-DM adaptada y las etapas planificadas del desarrollo de modelos, la evaluación y la creación de prototipos de software, con el objetivo final de fundamentar intervenciones efectivas y promover el éxito estudiantil en la FCBI-Unillanos.***

Palabras clave-- Minería de Datos Educativa (EDM), CRISP-DM, Predicción de Deserción Estudiantil, ETL.

I. Introducción

En el contexto global de la educación superior, la deserción estudiantil emerge como una problemática apremiante, acentuándose en la realidad latinoamericana y colombiana, donde las tasas de abandono pueden alcanzar niveles críticos, superando en ocasiones el 50% en ciertos programas académicos [1], [2]. Este fenómeno multifactorial, estudiado en diversas investigaciones, no solo impacta negativamente las trayectorias individuales de los estudiantes, truncando sus aspiraciones y limitando su movilidad social, sino que también genera significativos costos económicos y sociales para las instituciones educativas y para el desarrollo de la región en su conjunto [3], [4]. Las causas de la deserción son diversas y complejas, incluyendo factores socioeconómicos y familiares, dificultades de adaptación al ambiente universitario, problemas vocacionales y de elección de carrera, así como limitaciones en el apoyo institucional y el seguimiento académico [5], [6], [7]. En [5] identificaron factores críticos dentro de los cuales destacan la falta de asesoramiento, un entorno estudiantil poco propicio y un seguimiento académico insuficiente como elementos determinantes en la decisión de abandonar los estudios, resaltando la necesidad de intervenciones integrales que aborden las múltiples dimensiones de esta problemática.

Por otro lado, una investigación en Nicaragua [3] subraya la influencia de factores económicos, laborales y académicos en la deserción, mientras que un estudio en India [6] destaca la distancia a los centros de estudio y la percepción de cursos demandantes como barreras para la permanencia. En este sentido, resulta evidente que la deserción estudiantil no responde a una causa única, sino a una intrincada red de factores que interactúan de manera compleja y que requieren ser abordados desde una perspectiva holística y multidisciplinaria.

En este sentido, a pesar de la creciente conciencia sobre la magnitud e implicaciones de la deserción estudiantil, así como los esfuerzos investigativos por comprender sus causas y dinámicas; este fenómeno persiste como un desafío sustancial para las instituciones de educación superior en la actualidad. Bajo estas circunstancias, la minería de datos educativa (*EDM, por sus siglas en inglés*) se ha consolidado como un campo de investigación con un potencial significativo para transformar el análisis y la predicción de la deserción, gracias al avance de las tecnologías de la información y la creciente disponibilidad de datos.

Diversos estudios han explorado la aplicación de algoritmos de aprendizaje automático, como árboles de decisión [8], [9], [10], [11], [12], redes neuronales [13], [14], regresión lineal [15], [16] y otros métodos de clasificación [17], [18], [19]; para identificar patrones de riesgo, predecir el rendimiento académico y, en última instancia, actuar como mecanismo de predicción y en función de ello establecer estrategias para su prevención ante la deserción.

Estas herramientas analíticas ofrecen la capacidad de procesar grandes volúmenes de datos académicos, sociodemográficos e incluso comportamentales, revelando relaciones y factores predictivos que pueden pasar desapercibidos mediante métodos tradicionales. La globalización de estas tecnologías y su creciente accesibilidad plantean nuevas oportunidades para abordar el análisis de la deserción universitaria de manera más efectiva y personalizada, aunque también imponen la responsabilidad ética de utilizarlas de forma responsable y equitativa, garantizando la privacidad de los datos y evitando la reproducción de sesgos o discriminaciones [20].

En este marco, este trabajo se justifica por la necesidad de caracterizar la deserción estudiantil en la Facultad de Ciencias Básicas e Ingeniería (FCBI) de la Universidad de los Llanos (Unillanos) durante los años 2017 a 2022. Este estudio es resultado del proyecto de investigación institucionalizado donde se busca comprender y explicar este fenómeno en el contexto universitario. En este sentido se busca identificar patrones de deserción y descubrir información clave que exponga las correlaciones existentes entre diversos factores y los causales de deserción mediante la exploración de técnicas de minería de datos y aprendizaje automático, como árboles de decisión y redes neuronales [9], [10], [21].

Se parte de la hipótesis de que la aplicación de estas herramientas tecnológicas permitirá desarrollar modelos predictivos más robustos y personalizados, que a su vez faciliten a futuro el diseño e implementación de estrategias de intervención temprana para mitigar la deserción y promover la permanencia estudiantil en los programas de la FCBI.

Esta investigación preliminar plantea la oportunidad de aprovechar el potencial de las herramientas analíticas avanzadas en la comprensión y abordaje de un problema educativo. Se espera que los modelos y metodologías desarrolladas, basados en enfoques novedosos de minería de datos como CRISP-DM aplicado en este contexto [22], sean escalables a otras facultades e instituciones de educación superior. De esta forma, se busca contribuir en mejorar las oportunidades de formación y el futuro de los estudiantes, no solo en la región, sino potencialmente en todo el país.

II. Metodología

Este proyecto de investigación en progreso adopta como marco metodológico el proceso CRISP-DM (Cross-Industry Standard Process for Data Mining, *por sus siglas en ingles*), una metodología robusta y ampliamente reconocida para proyectos de minería de datos [21]. CRISP-DM se considera idónea para abordar el desafío de la predicción de la deserción estudiantil en la Facultad de Ciencias Básicas e Ingeniería (FCBI) de la Unillanos, debido a su estructura en fases, su enfoque iterativo y su flexibilidad para adaptarse a diferentes contextos [20].

En particular, su aplicabilidad en el contexto de la deserción estudiantil ha sido demostrada en investigaciones previas. Se ha aplicado CRISP-DM para analizar la deserción académica y proponer un método basado en Descubrimiento de Conocimiento en Bases de Datos (KDD, *por sus siglas en ingles*) para identificar patrones de comportamiento estudiantil [23]. Además, la relevancia de metodologías de inteligencia computacional en el contexto colombiano, donde se enmarca esta investigación, se evidencia en trabajos como los presentados en [24]. La estructura de CRISP-DM, con sus fases bien definidas, permite abordar de manera sistemática la complejidad inherente a la predicción de la deserción, facilitando la extracción de conocimiento valioso a partir de los datos disponibles.

El diseño metodológico se plantea en función de la metodología de CRISP-DM, que integra explícitamente el desarrollo de software como un componente central, permitiendo no solo la caracterización, sino también la implementación de un prototipo de sistema para la detección temprana de estudiantes en riesgo. La Figura 1, presenta un diagrama de flujo que visualiza las fases de esta metodología CRISP-DM que se implementará en el proyecto, detallando las tareas principales en cada etapa y resaltando su naturaleza iterativa en el contexto específico de la FCBI-Unillanos.

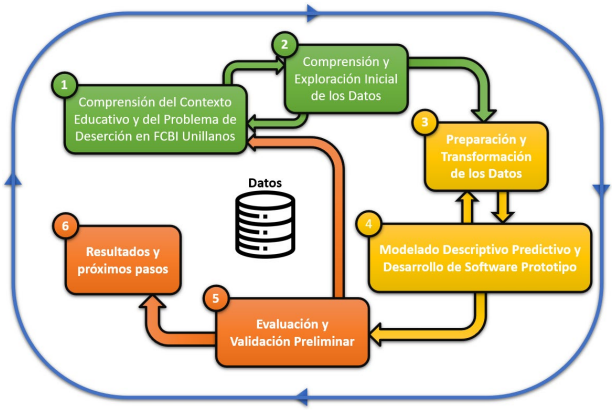

Fig. 1 Diagrama de Flujo de la Metodología CRISP-DM Adaptada para la Predicción de Deserción Estudiantil en la FCBI-Unillanos.

III. Resultados Preliminares

Como un resultado preliminar fundamental de este Trabajo en Progreso, se presenta el diseño de una metodología adaptada basada en CRISP-DM, visualizada en el diagrama de flujo de la Figura 1. Este diagrama resume el enfoque metodológico que guía la investigación, orientada a un doble objetivo: identificar patrones de deserción estudiantil en la FCBI-Unillanos con mayor precisión y descubrir información clave sobre sus causales subyacentes. La metodología, inspirada en CRISP-DM 1.0, ha sido cuidadosamente adaptada al contexto específico de la FCBI-Unillanos y al objetivo de desarrollar un sistema predictivo, integrando el desarrollo de software como un componente esencial. La claridad y estructura de esta metodología adaptada, representada visualmente en el diagrama de flujo, constituyen un avance significativo en la planificación y ejecución del proyecto.

En estrecha sinergia con este planteamiento metodológico, otro resultado preliminar relevante es la revisión de literatura realizada en el campo de la Minería de Datos Educativa (EDM) y la predicción de la deserción estudiantil. Esta revisión bibliográfica ha sido importante para fundamentar teóricamente la investigación, identificar variables relevantes utilizadas en

estudios previos, y seleccionar algoritmos de modelado pertinentes para el contexto de la FCBI-Unillanos.

Actualmente, la investigación se encuentra en la fase de Preparación de Datos. El trabajo en progreso destaca que los esfuerzos se han concentrado en la comprensión exhaustiva de los datos disponibles, provenientes del sistema de información institucional unillanos (SIAU) y archivos proporcionados por la oficina de sistemas de la universidad. Paralelamente, se avanza en el diseño e implementación del proceso de Extracción, Transformación y Carga (ETL, *por sus siglas en ingles*), con el objetivo de asegurar la calidad, coherencia y adecuación de los datos para las fases de modelado predictivo. Las tareas en curso incluyen la definición de estrategias para la limpieza de datos, la codificación de variables categóricas relevantes y la selección de características que se anticipan como más informativas para la predicción de la deserción y la identificación de sus factores causales.

AGRADECIMIENTO/RECONOCIMIENTO

La Dirección General de Investigación de la Universidad de Los Llanos, Colombia, apoyó y financió este trabajo bajo el proyecto de investigación titulado "Caracterización de la deserción estudiantil en la Facultad de Ciencias Básicas e Ingeniería de la Unillanos. 2017-2022", código: C10-F02-006-2024.

Adicionalmente, este proyecto fue posible gracias al apoyo brindado por la Universidad de los Llanos a través de la "Convocatoria para el Fortalecimiento de las Capacidades Científicas, Tecnológicas y de Innovación Articuladas con la Agenda Institucional de Investigaciones: Jóvenes Investigadores Corte II 2024". Expresamos nuestro sincero agradecimiento a la Universidad de los Llanos por esta valiosa oportunidad que brinda a jóvenes investigadores egresados de la institución para fortalecer sus capacidades investigativas e innovadoras, haciendo posible la realización de este trabajo en progreso.

REFERENCIAS

[1] S. C. Guerrero and R. L. Espejo, "University dropout: comparative study between Colombia and Spain from a gender perspective," *Formacion Universitaria*, vol. 17, no. 2, pp. 101–112, 2024, doi: 10.4067/S0718-50062024000200101.

[2] L. Cabrera, J. T. Bethencourt, P. Alvarez Pérez, and M. González Afonso, "El problema del abandono de los estudios universitarios," *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, vol. 12, no. 2, Oct. 2014, doi: 10.7203/RELIEVE.12.2.4226.

[3] L. R. Calderón, "DESERCIÓN EN LA EDUCACIÓN SUPERIOR RECINTO LAS MINAS. PERÍODO 2001-2007," *Ciencia e Interculturalidad*, vol. 4, no. 1, pp. 30–46, Mar. 2011, doi: 10.5377/RCI.V4I1.288.

[4] M. Duarte and M. Isabel, "Análisis de la deserción universitaria en Colombia a partir de la revisión bibliográfica 2015 - 2021," 2022. Accessed: Feb. 15, 2025. [Online]. Available: https://hdl.handle.net/20.500.12495/9647

[5] A. B. Urbina-Nájera, J. C. Camino-Hampshire, and R. Cruz Barbosa, "Deserción escolar universitaria: Patrones para prevenirla aplicando minería de datos educativa," *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, vol. 26, no. 1, pp. 1–19, Oct. 2020, doi: 10.7203/RELIEVE.26.1.16061.

[6] B. I. Fozdar, L. S. Kumar, and S. Kannan, "Study of the Factors Responsible for the Dropouts from the BSc Programme of Indira Gandhi National Open University," *The International Review of Research in Open and Distributed Learning*, vol. 7, no. 3, Dec. 2006, doi: 10.19173/IRRODL.V7I3.291.

[7] S. C. Guerrero, "Deserción universitaria: Políticas y vivencias de docentes y estudiantes en una universidad colombiana (2010-2017)," *Editorial UPTC*, Jul. 2023, doi: 10.19053/9789586607735.

[8] S. Sivakumar, S. Venkataraman, and R. Selvaraj, "Predictive modeling of student dropout indicators in educational data mining using improved decision tree," *Indian J Sci Technol*, vol. 9, no. 4, pp. 1–5, 2016, doi: 10.17485/IJST/2016/V9I4/87032.

[9] M. A. Al-Barrak and M. Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study," *International Journal of Information and Education Technology*, vol. 6, no. 7, pp. 528–533, 2016, doi: 10.7763/IJIET.2016.V6.745.

[10] B. Gupta, A. Rawat, A. Jain, A. Arora, and N. Dhami, "Analysis of Various Decision Tree Algorithms for Classification in Data Mining," *Int J Comput Appl*, vol. 163, no. 8, pp. 15–19, Apr. 2017, doi: 10.5120/IJCA2017913660.

[11] G. S. Abu-Oda and A. M. El-Halees, "Data Mining in Higher Education : University Student Dropout Case Study," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 1, pp. 15–27, Jan. 2015, doi: 10.5121/IJDKP.2015.5102.

[12] S. Himani and K. Sunil, "A Survey on Decision Tree Algorithms of Classification in Data Mining," *International Journal of Science and Research (IJSR)*, vol. 5, no. 4, pp. 2094–2097, Apr. 2016, doi: 10.21275/v5i4.NOV162954.

[13] E. Yukselturk, S. Ozekes, and Y. K. Türel, "Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program," *European Journal of Open, Distance and E-Learning*, vol. 17, no. 1, pp. 118–133, Jul. 2014, doi: 10.2478/EURODL-2014-0008.

[14] M. Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education," *IEEE Access*, vol. 4, pp. 2379–2387, 2016, doi: 10.1109/ACCESS.2016.2568756.

[15] S. J. H. Yang, O. H. T. Lu, A. Y. Q. Huang, J. C. H. Huang, H. Ogata, and A. J. Q. Lin, "Predicting students' academic performance using multiple linear regression and principal component analysis," *Journal of Information Processing*, vol. 26, pp. 170–176, Jan. 2018, doi: 10.2197/IPSJJIP.26.170.

[16] E. Yamao, L. Celi Saavedra, R. Campos Pérez, and V. de J. Huancas Hurtado, "Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo en una universidad peruana," *Campus*, vol. 23, no. 26, pp. 151–160, Dec. 2018, doi: 10.24265/CAMPUS.2018.V23N26.05.

[17] J. Sayali and H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques," *International Journal of Science and Research (IJSR)*, vol. 5, no. 1, pp. 1842–1845, Jan. 2016, doi: 10.21275/V5I1.NOV153131.

[18] W. Chen, X. Xie, J. Peng, J. Wang, Z. Duan, and H. Hong, "GIS-based landslide susceptibility modelling: a comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models," *Geomatics, Natural Hazards and Risk*, vol. 8, no. 2, pp. 950–973, Dec. 2017, doi: 10.1080/19475705.2017.1289250.

[19] R. I. Estrada-Danell, R. A. Zamarripa-Franco, P. G. Zúñiga-Garay, and I. Martínez-Trejo, "Aportaciones desde la minería de datos al proceso de captación de matrícula en Instituciones de Educación Superior particulares.," *Revista Electrónica Educare*, vol. 20, no. 3, p. 1, Sep. 2016, doi: 10.15359/REE.20-3.11.

[20] J. L. Aznarte, "Ethical considerations on the use of massive data-based technologies in UNED," *RIED-Revista Iberoamericana de*

Educacion a Distancia, vol. 23, no. 2, pp. 237–252, Jul. 2020, doi: 10.5944/RIED.23.2.26590.

[21] F. Seoane, E. Fernández-Díaz, and N. López, "Dropout Prediction in Higher Education with Artificial Intelligence: A Convolutional Neural Network Approach," *International Journal of Emerging Technologies in Learning (iJET)*, 2020, [Online]. Available: https://eric.ed.gov/?id=EJ1277899

[22] A. M. Shimaoka, R. C. Ferreira, and A. Goldman, "The evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks," *SBC Reviews on Computer Science*, vol. 4, no. 1, pp. 28–43, Oct. 2024, doi: 10.5753/REVIEWS.2024.3757.

[23] L. F. Castro R., E. Espitia P., and A. F. Montilla, "Applying CRISP-DM in a KDD process for the analysis of student attrition," in *Communications in Computer and Information Science*, Springer Verlag, 2018, pp. 386–401. doi: 10.1007/978-3-319-98998-3_30.

[24] Boris Perez, Camilo Castellanos, and Dario Correal, *Applying Data Mining Techniques to Predict Student Dropout: A Case Study*. IEEE 1st Colombian Conference on Applications in Computational Intelligence, 2018.