Early prediction of students at risk of low academic performance through data mining techniques and machine learning

Gerzel Stella Maris[1] 0009-0006-0331-1842
Brollo Magdalena Elisabet[1] 0009-0009-9855-7301 Carlos Emanuel Aguirre[1] 0009-0009-9662-956x
Paszco Gustavo Ariel[1] 0009-0002-4794-7296

[1] Universidad Nacional del Chaco Austral, Chaco, Argentina
stellagerzel@uncaus.edu.ar, magdalenabrollo@uncaus.edu.ar,
aguirreemanuel@uncaus.edu.ar, paszcogustavo@uncaus.edu.ar

Abstract. Low academic performance and early dropout rates during the initial years of higher education constitute significant challenges for universities, given their direct impact on student trajectories and institutional retention indicators. This paper presents an approach grounded in machine learning techniques aimed at the early identification of academically at-risk students within the Systems Engineering program at the Universidad Nacional del Chaco Austral (UNCAus). The study utilizes a dataset comprising 978 academic records from the courses Algorithms and Data Structures and Systems and Organizations, spanning the 2024 and 2025 cohorts. This dataset integrates socioeconomic, familial, employment, technological, and motivational variables collected through institutional surveys. Three predictive classification tasks were defined: Overall academic risk (passing versus failing or dropping out). Academic failure. Absenteeism or dropout. For each scenario, Logistic Regression, Decision Trees, Random Forest, and XGBoost models were evaluated, employing stratified cross-validation alongside various class-balancing strategies. The optimal results were achieved using Random Forest for predicting overall academic risk and failure, whereas XGBoost proved superior in identifying students at risk of dropping out. The models yielded ROC-AUC values of up to 0.69 for dropout prediction and successfully identified critical variables associated with academic performance, including family characteristics, socioeconomic conditions, access to technological resources, and motivational factors. In conclusion, these findings demonstrate the potential of these techniques as decision-support tools for the early detection of vulnerable students and the implementation of preemptive pedagogical interventions.

Keywords: educational data science, early prediction, academic performance, data mining, machine learning.

Predicción temprana de estudiantes en riesgo de bajo rendimiento académico mediante técnicas de minería de datos y aprendizaje automático

Resumen. El bajo rendimiento académico y el abandono durante los primeros años universitarios constituyen problemáticas relevantes para las instituciones de educación superior debido a su impacto sobre las trayectorias estudiantiles y los indicadores de permanencia. Este trabajo presenta una propuesta basada en técnicas de aprendizaje automático para la identificación temprana de estudiantes en riesgo académico en la carrera de Ingeniería en Sistemas de la Universidad Nacional del Chaco Austral (UNCAus). El estudio utiliza información correspondiente a 978 registros académicos de las asignaturas Algoritmos y Estructuras de Datos y Sistemas y Organizaciones de las cohortes 2024 y 2025. El conjunto de datos integra variables socioeconómicas, familiares, laborales, tecnológicas y motivacionales obtenidas mediante relevamientos institucionales. Se definieron tres problemas de clasificación: riesgo académico global (aprobación versus desaprobación o abandono), desaprobación académica y ausencia o abandono. Para cada caso se evaluaron modelos de Regresión Logística, Árboles de Decisión, Random Forest y XGBoost, utilizando validación cruzada estratificada y distintas estrategias de balanceo de clases. Los mejores resultados se obtuvieron mediante Random Forest para la predicción de riesgo académico global y desaprobación, y mediante XGBoost para la identificación de estudiantes con riesgo de abandono. Los modelos alcanzaron valores de ROC-AUC de hasta 0.69 en la predicción de abandono. Los resultados muestran el potencial de estas técnicas como herramientas de apoyo para la detección temprana de estudiantes vulnerables y la implementación de intervenciones pedagógicas preventivas.

Palabras clave: ciencia de datos educacional, predicción temprana, rendimiento académico, minería de datos, aprendizaje automático.

1 Introducción

El riesgo de bajo rendimiento académico y la deserción en los primeros años de la educación superior constituyen desafíos estructurales que afectan la equidad educativa, la eficiencia institucional y el desarrollo socioeconómico. Diversos estudios señalan que el desempeño en el primer año, especialmente en asignaturas introductorias de programación, es un factor determinante en la persistencia estudiantil (Bernardo et al., 2016; Bennedsen & Caspersen, 2019). Asimismo, la falta de mecanismos de detección temprana suele derivar en intervenciones tardías e ineficaces (Seymour & Hunter, 2019).

En este contexto, la creciente disponibilidad de datos educativos ha impulsado el desarrollo de la Minería de Datos Educativa (EDM) y el Learning Analytics (LA), disciplinas orientadas a la generación de conocimiento a partir de datos para mejorar los procesos de enseñanza y aprendizaje (Romero & Ventura, 2020). La literatura

reciente destaca el uso de técnicas de aprendizaje automático para modelar la complejidad del rendimiento académico, superando enfoques estadísticos tradicionales (Kotsiantis, 2012). En particular, estudios previos han demostrado la efectividad de estas técnicas para la predicción del abandono y del desempeño académico a partir de variables académicas y contextuales (Cortez & Silva, 2008; Dekker et al., 2009), así como la necesidad de integrar dimensiones socioeconómicas y actitudinales en el análisis ((La Red Martínez et al., 2016); Albreiki et al., 2021; Baashar et al., 2022).

En este marco, el presente trabajo propone un sistema de predicción temprana del riesgo de bajo rendimiento académico en estudiantes de Ingeniería en Sistemas de la Universidad Nacional del Chaco Austral (UNCAus), basado en un repositorio que integra variables académicas, socioeconómicas, laborales y actitudinales. Las principales contribuciones de este trabajo son la construcción de un dataset contextualizado y el desarrollo de modelos de aprendizaje automático orientados a la detección temprana de estudiantes en riesgo académico.

2 Metodología

La investigación adopta un enfoque cuantitativo basado en técnicas de Minería de Datos Educativa (Educational Data Mining, EDM) y aprendizaje automático para la identificación temprana de estudiantes en riesgo académico. El objetivo consiste en desarrollar modelos predictivos capaces de anticipar situaciones de desaprobación y abandono en los primeros años de la carrera de Ingeniería en Sistemas de la Universidad Nacional del Chaco Austral (UNCAus).

2.1 Contexto y Conjunto de Datos

El estudio se desarrolló sobre información correspondiente a las asignaturas Algoritmos y Estructuras de Datos y Sistemas y Organizaciones durante las cohortes 2024 y 2025. Ambas materias presentan históricamente elevados índices de desaprobación y constituyen asignaturas críticas dentro del primer año de la carrera. Se construyó un repositorio compuesto por 978 registros de estudiantes, que integra variables provenientes del relevamiento institucional mediante el sistema académico SIU-Guarani. Las variables consideradas abarcan dimensiones: socioeconómicas; familiares; laborales; tecnológicas; motivacionales y académicas.

Entre las variables incluidas se encuentran cobertura de salud, tipo de vivienda, situación laboral, convivencia familiar, acceso a recursos tecnológicos, práctica deportiva, recepción de becas y motivaciones para la elección de la carrera.

Previamente al modelado se realizaron procesos de depuración, codificación, normalización y tratamiento de valores faltantes.

2.2 Definición de los Problemas de Clasificación

Con el propósito de analizar diferentes manifestaciones del riesgo académico se definieron tres problemas de clasificación binaria: Problema 1: Riesgo académico global (A vs R+U). Clase 0: estudiantes aprobados (A). Clase 1: estudiantes desaprobados o ausentes/abandonaron (R+U). Problema 2: Riesgo de desaprobación (A vs R). Clase 0: estudiantes aprobados (A). Clase 1: estudiantes desaprobados (R).

Problema 3: Riesgo de ausencia o abandono (A vs U). Clase 0: estudiantes aprobados (A). Clase 1: estudiantes ausentes o abandonaron la asignatura (U).

Esta estrategia permite estudiar separadamente los factores asociados al bajo rendimiento académico y aquellos vinculados específicamente al abandono.

2.3 Modelos Utilizados

Se evaluaron cuatro algoritmos supervisados ampliamente utilizados en problemas de clasificación: Regresión Logística, Árbol de Decisión, Random Forest y XGBoost. La selección de estos modelos responde a su capacidad para manejar conjuntos de datos heterogéneos, variables categóricas y relaciones no lineales entre atributos.

Ante el desbalance observado entre las clases, se analizaron tres estrategias de tratamiento: ponderación de clases, sobremuestreo sintético mediante SMOTE y submuestreo aleatorio (undersampling). Cada estrategia fue evaluada independientemente para determinar su impacto sobre el desempeño predictivo.

2.5 Validación

La validación de los modelos se realizó mediante validación cruzada estratificada, utilizando como métrica principal el área bajo la curva ROC (ROC-AUC). Adicionalmente se calcularon las métricas de: accuracy, precisión, recall y F1-score. Posteriormente se analizaron distintos umbrales de decisión para estudiar el compromiso entre sensibilidad y cantidad de falsos positivos.

3 Resultados Preliminares

Se realizó la evaluación de los cuatro algoritmos de clasificación, combinados con las tres estrategias de balanceo de clases, en los que los resultados de los algoritmos de Regresión Logística y Árbol de Decisión han sido inferiores a los otros, por lo que no se los incluyó en el proceso de calibración (Threshold). De los resultados del ajuste del umbral de sensibilidad se pudo ver que la combinación Random Forest con SMOTE obtuvo los mejores valores de ROC-AUC, 0.600 y 0.590 respectivamente, para los problemas de riesgo académico global y desaprobación. En cambio, XGBoost con ponderación de clases fue la mejor configuración alcanzando un ROC-AUC de 0.694, para la predicción de ausencia o abandono. La Tabla 1 resume los mejores resultados obtenidos para cada problema de clasificación planteado.

Tabla 1: Mejores algoritmos obtenidos para cada problema de clasificación

Problema	A vs R+U	A vs R	A vs U
Modelo	Random Forest	Random Forest	XGBoost
Estrategia	SMOTE	SMOTE	class_weight
Threshold	0.30	0.30	0.40
ROC-AUC	0.600	0.590	0.692
Accuracy	0.712	0.621	0.645
Precisión	0.717	0.622	0.591
Recall	0.984	0.982	0.798
F1-score	0.830	0.762	0.679

El umbral 0.40 fue el valor óptimo identificado para el modelo del problema de abandono durante el proceso de calibración. Como alternativa institucional puede emplearse 0.30 para incrementar la sensibilidad, aunque con un aumento de falsos positivos.

En la Tabla 2 se muestra la matriz de confusión de los modelos seleccionados. Donde TN = casos negativos correctamente clasificados. FP = falsos positivos. FN = falsos negativos. TP = casos positivos correctamente clasificados. En A vs R+U, la clase positiva corresponde a riesgo académico global; en A vs R, a desaprobación; y en A vs U, a abandono/ausencia.

Tabla 2: Matriz de confusión de los modelos seleccionados

Problema	Modelo	TN	FP	FN	TP
A vs R+U	Random Forest y SMOTE	8	271	11	688
A vs R	Random Forest y SMOTE	10	269	8	443
A vs U	XGBoost y class_weight	142	137	50	198

El análisis de las matrices de confusión muestra que los modelos seleccionados privilegian la detección temprana de estudiantes en situación de riesgo, minimizando la cantidad de falsos negativos. En los problemas de riesgo académico global (A vs R+U) y desaprobación (A vs R) se obtuvieron 11 y 8 falsos negativos respectivamente, lo que evidencia una elevada sensibilidad para identificar estudiantes vulnerables. Sin embargo, esta estrategia incrementa la cantidad de falsos positivos. Por su parte, el modelo enfocado en la deserción (A vs U), presentó un comportamiento más equilibrado, reduciendo significativamente las falsas alarmas (FP = 137) a costa de incrementar el número de falsos negativos (FN = 50).

4 Discusión y Conclusiones

El trabajo presentó un enfoque basado en técnicas de aprendizaje automático para la identificación temprana de estudiantes en riesgo académico en asignaturas críticas del primer año de Ingeniería en Sistemas de la UNCAus.

Los resultados preliminares muestran que Random Forest constituye la alternativa más adecuada para la predicción de riesgo académico global y desaprobación, mientras que XGBoost obtuvo el mejor desempeño para la identificación de estudiantes con riesgo de abandono.

La utilización de distintos umbrales de decisión permitió analizar el compromiso existente entre sensibilidad y precisión. Desde una perspectiva institucional, la priorización del recall resulta razonable debido al elevado costo asociado a no detectar estudiantes en riesgo. Sin embargo, esta decisión implica un incremento en la cantidad de falsos positivos, aspecto que debe ser considerado durante la implementación práctica del sistema.

Debido a que los modelos incorporan variables socioeconómicas, familiares y geográficas, existe el riesgo potencial de reproducir desigualdades presentes en los datos históricos. Por esta razón, los resultados obtenidos no deben utilizarse para la toma automática de decisiones académicas ni para la clasificación permanente de estudiantes. Su finalidad consiste exclusivamente en asistir a docentes y responsables

institucionales en la identificación temprana de situaciones de vulnerabilidad académica que requieran acompañamiento pedagógico.

La desaprobación y el riesgo académico global mostraron desempeños moderados, se espera que, al incluir otras variables relacionadas con el desempeño académico, por ejemplo: notas obtenidas en los trabajos y parciales, tiempo dedicado al estudio, etc. los resultados sean más favorables. La predicción de abandono alcanzó los mejores resultados, sugiriendo la existencia de patrones más claramente identificables asociados a la desvinculación temprana.

Como trabajo futuro se propone ampliar el conjunto de datos, incorporar nuevas cohortes académicas y variables relacionadas con el desempeño.

Agradecimientos: Uso de Data Mining y producción de materiales multimedia para el mejoramiento de la Enseñanza – Aprendizaje de algunos temas de Álgebra y de Lógica y Matemática Computacional. Proyecto 22F017 - UNNE

Referencias

Albreiki, B., Zaki, N., y Alashwal, H. (2021). *A systematic literature review of predicting student performance using machine learning techniques*. IEEE Access, 9, 41803-41829.

Baashar, Y., Alkawsi, G., Ali, N. A., Alhussian, H., y Bahbouh, H. T. (2021). *Predicting student's performance using machine learning methods: A systematic literature review.* 2021 International Conference on Computer & Information Sciences (ICCOINS), 1–6. https://doi.org/10.1109/ICCOINS49675.2021.9515151

Bennedsen, J., y Caspersen, M. E. (2019). *Failure rates in introductory programming*. ACM SIGCSE Bulletin, 51(2), 32-38. https://doi.org/10.1145/3328334.3328343.

Bernardo, A. B., Esteban, M., Fernández, E., Cervero, A., Tuero, E., y Solano, P. (2016). *Variables predictoras del abandono universitario: el papel de la orientación profesional y la interacción con el profesor*. Psicothema, 28(1), 46-52.

Cortez, P., y Silva, A. (2008). *Using data mining to predict secondary school student performance.* In Proceedings of the 5th Future Business Technology Conference (pp. 5–12).

Dekker, G. W., Pechenizkiy, M., y Vleeshouwers, J. M. (2009). *Predicting students drop out: A case study.* Proceedings of the International Conference on Educational Data Mining, 41–50.

Kotsiantis, S. B. (2012). *Use of machine learning techniques for educational data mining*. Journal of Educational Computing Research, 7(2), 469-485.

La Red Martínez, D. L., Karanik, M., Giovannini, M. E., y Scappini, R. (2016). *Towards a predictive model of academic performance using data mining in the UTN-FRRe*. Journal of Systemics, Cybernetics and Informatics, 14(2), 36–41.

Romero, C., y Ventura, S. (2020). *Educational data mining and learning analytics: An updated survey*. WIREs Data Mining and Knowledge Discovery, 10(3), e1355.

Seymour, E., y Hunter, A. B. (Eds.). (2020). *Talking about Leaving Revisited: Persistence, Relocation, and Loss in Undergraduate STEM Education*. Springer Nature. https://doi.org/10.1007/978-3-030-25304-2

Durante la preparación de este trabajo, los autores utilizamos las herramientas de inteligencia artificial ChatGPT y Gemini con el fin de mejorar la redacción y traducción al idioma inglés. Tras utilizar estas herramientas, hemos revisado y editado el contenido según fuera necesario y asumimos plena responsabilidad por el contenido de la publicación.