Predicción de Rendimiento Estudiantil

Santiago Martínez Beltrán
Facultad de Ciencias Naturales e Ingeniería
Universidad Jorge Tadeo Lozano
Bogotá, Colombia
santiago.martinezb@utadeo.edu.co

Sergio Daniel Aza Ocampo
Facultad de Ciencias Naturales e Ingeniería
Universidad Jorge Tadeo Lozano
Bogotá, Colombia
Sergiod.azaocampo@utadeo.edu.co

Julián Santiago Hernández González
Facultad de Ciencias Naturales e Ingeniería
Universidad Jorge Tadeo Lozano
Bogotá, Colombia
Julians.hernandezg@uatdeo.edu.co

Curso: Inteligencia Artificial

5 de octubre de 2025

Resumen

Este proyecto propone el desarrollo de un modelo de aprendizaje automático capaz de predecir el rendimiento académico de los estudiantes universitarios a partir de variables socioeducativas y de comportamiento. El objetivo es identificar factores que influyen en el desempeño y anticipar posibles riesgos de bajo rendimiento, con el fin de apoyar estrategias de acompañamiento académico en la Universidad Jorge Tadeo Lozano.
Se utilizará el dataset *Student Performance* del repositorio UCI Machine Learning [1], el cual contiene información de 649 estudiantes de educación media en Portugal.
El enfoque de IA corresponde a una tarea de clasificación con modelos como Árboles de Decisión y Regresión Logística [7][8][9].
Se espera lograr una precisión superior al 80% y comprobar que variables como las horas de estudio y las ausencias tienen alta correlación con el rendimiento final [4][6].

Problema Local y Motivación

En el contexto universitario bogotano, el rendimiento académico de los estudiantes es un factor crítico que influye en la deserción y en el éxito profesional. En la Universidad Jorge Tadeo Lozano, al igual que en muchas instituciones, los docentes suelen carecer de herramientas predictivas que les permitan identificar tempranamente a estudiantes en riesgo de bajo desempeño.
El proyecto busca aplicar inteligencia artificial para apoyar la gestión académica mediante la predicción del rendimiento a partir de información disponible, como hábitos de estudio, notas parciales y asistencia [2][3][5].
Esto permitiría orientar intervenciones personalizadas y promover la permanencia estudiantil, impactando positivamente la calidad educativa y la eficiencia institucional [4].

Dataset

Se empleará el *Student Performance Dataset* del repositorio **UCI Machine Learning** [1]. Este conjunto contiene 649 registros de estudiantes y 33 variables, incluyendo factores demográficos, sociales y académicos.
Las variables más relevantes para el modelo son: studytime, failures, absences, G1, G2 y G3 (notas).
El dataset está disponible bajo licencia pública para investigación educativa y es representativo por su diversidad de características, lo cual lo hace adecuado para modelar la predicción del rendimiento académico [1][4].
Los datos se encuentran en formato CSV, lo que facilita su manipulación y procesamiento mediante bibliotecas de Python como *pandas* y *scikit-learn* [10]

Tarea de IA y Algoritmo(s)

La tarea corresponde a un problema de **clasificación supervisada** en datos tabulares [5][6].
El objetivo del modelo es predecir si un estudiante tendrá un rendimiento alto o bajo según sus características previas.
Se emplearán los algoritmos **Árbol de Decisión** y **Regresión Logística**, por su facilidad de interpretación

y eficiencia en datasets pequeños [7][8][9]. El Árbol de Decisión permitirá visualizar los factores determinantes, mientras que la Regresión Logística ofrecerá una referencia estadística como línea base [7][8].

Metodología y Evaluación

El proceso inicia con la limpieza del dataset y la codificación de variables categóricas mediante *Label Encoding* o *One-Hot Encoding* [9]. Posteriormente, se dividirán los datos en conjuntos de entrenamiento (80%) y prueba (20%) [5]. Se evaluará el modelo utilizando métricas como *Accuracy*, *Precision*, *Recall* y *F1-Score*, además de una matriz de confusión para interpretar los errores [9][10].

Como línea base, se compararán los resultados con un clasificador aleatorio, verificando si la combinación de árboles de decisión y regresión logística mejora el rendimiento de la predicción [4][6].

Resultados esperados, ética y cronograma

Se espera que el modelo prediga con alta exactitud el rendimiento final y determine qué factores influyen más en el desempeño académico [4][6]. Los resultados permitirán crear estrategias preventivas en la universidad, contribuyendo a mejorar la retención y la toma de decisiones institucionales [2][3].

Desde el punto de vista ético, se garantizará la anonimización de los datos y el uso exclusivo con fines académicos, evitando sesgos por género o condición socioeconómica [3][8].

Cronograma propuesto:

- **Semana 1:** Revisión del problema, recolección del dataset y exploración inicial de variables.
- **Semana 2:** Limpieza y preprocesamiento de los datos (codificación, normalización y selección de variables).
- **Semana 3:** Entrenamiento de los modelos (Árbol de Decisión y Regresión Logística) y ajuste de parámetros.
- **Semana 4:** Evaluación del desempeño de los modelos, análisis de métricas y visualización de resultados.
- **Semana 5:** Redacción final del informe, revisión ética, corrección de estilo y entrega del paper.

Roles de Equipo

Rol	Integrante	Responsabilidad
Análisis de datos y preprocesamiento	Persona 1	Obtención, limpieza y análisis del Dataset
Modelado y métricas	Persona 2	Entrenamiento y evaluación de modelos
Redacción y ética	Persona 3	Elaboración del informe, revisión ética y cronograma

Enlace Repositorio de Github:

https://github.com/Santiageoff/ProyectoIA.git

Referencias

[1] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," University of Minho, Portugal, UCI Machine Learning Repository, 2008. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Student+Performance.

[2] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 3, e1355, 2020, doi: 10.1002/widm.1355.

[3] V. Kumar and A. Chadha, "An empirical study of the applications of data mining techniques in higher education," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 3, no. 2, pp. 80–84, 2012, doi: 10.14569/IJACSA.2012.030214.

[4] A. N. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student performance using data mining techniques," Procedia Computer Science, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.111.

[5] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," Applied Artificial Intelligence, vol. 18, no. 5, pp. 411–426, 2004, doi: 10.1080/08839510490442058.

[6] S. Huang, N. Fang, and N. Chen, "Predicting student academic performance with educational data mining," Frontiers in Education Technology, vol. 1, no. 1, pp. 1–8, 2018, doi: 10.22158/fet.v1n1p1.

[7] M. Kumar and A. Singh, "Predicting student performance using decision tree algorithms," International Journal of Computer Applications, vol. 175, no. 5, pp. 22–26, 2017, doi: 10.5120/ijca2017914605.

[8] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, 4th ed. Birmingham, UK: Packt Publishing, 2022. [Online]. Available: https://www.packtpub.com/product/python-machine-learning-fourth-edition/9781801819312.

[9] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow*, 3rd ed. Sebastopol, CA: O'Reilly Media, 2022. [Online]. Available: https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html.