Optimización Multi-Objetivo para Clasificación de Rendimiento Estudiantil con Minería de Datos

Angello Marcelo Zamora Valencia
Facultad de Ingeniería Estadística e Informática
Universidad Nacional del Altiplano
Puno, Perú

Resumen

Este estudio presenta la aplicación de algoritmos de optimización multi-objetivo para la clasificación automática de estudiantes en perfiles académicos basados en sus hábitos de estudio y rendimiento. Se implementaron y compararon tres algoritmos evolutivos: (NSGA-II), (MOPSO) y (MOEA/D) utilizando un dataset de 1000 estudiantes con 16 variables comportamentales y académicas. Los objetivos optimizados simultáneamente fueron: minimización de la tasa de error de clasificación, reducción del número de características, minimización de la complejidad computacional y reducción de violaciones de equidad entre grupos. MOEA/D alcanzó la mejor precisión con 81.6 % ± 2.1 % utilizando 7.8 características promedio, mientras que NSGA-II logró 79.2 % ± 2.8 % con mayor eficiencia computacional. Se identificaron cuatro perfiles estudiantiles distintos: Alto Rendimiento (22 %), Equilibrado (58 %), En Riesgo (18 %) y Social (12 %). Los resultados demuestran la superioridad de los enfoques multi-objetivo sobre métodos tradicionales de objetivo único, proporcionando soluciones balanceadas entre precisión, interpretabilidad y eficiencia computacional para sistemas de apoyo académico.

Palabras clave: Optimización multi-objetivo, NSGA-II, MOPSO, MOEA/D, minería de datos educacionales, clasificación estudiantil, algoritmos evolutivos.

1. Introducción

La minería de datos educacionales ha emergido como una disciplina fundamental para comprender los patrones complejos que determinan el éxito académico estudiantil. La naturaleza multidimensional del rendimiento educativo, que abarca factores cognitivos, comportamentales, sociales y ambientales, requiere enfoques analíticos sofisticados capaces de modelar múltiples objetivos simultáneamente [1].

Los métodos tradicionales de clasificación estudiantil se han centrado en la optimización de un único objetivo, típicamente la precisión de predicción, ignorando otros aspectos críticos como la interpretabilidad del modelo, la eficiencia computacional y la equidad entre diferentes grupos demográficos. Esta limitación ha motivado la adopción de algoritmos de optimización multi-objetivo (MOO) que pueden equilibrar múltiples criterios competidores simultáneamente [2].

Los algoritmos evolutivos multi-objetivo han demostrado particular eficacia en dominios educacionales debido a su capacidad para explorar eficientemente espacios de búsqueda complejos y generar conjuntos de soluciones Pareto-óptimas. Entre estos, NSGA-II se ha establecido como referencia debido a su mecanismo de clasificación no dominada rápida y preservación de diversidad mediante distancia de hacinamiento [3]. MOPSO ofrece ventajas en términos de velocidad de convergencia, mientras que MOEA/D sobresale en problemas de muchos objetivos mediante su estrategia de descomposición [4].

La pandemia COVID-19 ha acelerado la generación de datos educacionales digitales, creando oportunidades sin precedentes para el análisis predictivo del rendimiento estudiantil. Sin embargo, esta abundancia de datos también presenta desafíos en términos de dimensionalidad, interpretabilidad y equidad algorítmica [5].

Este trabajo contribuye al estado del arte mediante: (1) la primera comparación exhaustiva de algoritmos MOO aplicados a clasificación de perfiles estudiantiles usando datos reales, (2) la formulación de un problema multi-objetivo que balancea precisión, interpretabilidad, eficiencia y equidad, (3) la identificación de perfiles estudiantiles distintos basados en análisis de optimización Pareto, y (4) la validación empírica usando un dataset de 1000 estudiantes con 16 variables comportamentales y académicas.

2. Métodos

El problema de clasificación multi-objetivo se formuló como la minimización simultánea del vector de objetivos $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x}), f_4(\mathbf{x})]^T$, donde f_1 representa la tasa de error de clasificación, f_2 el número de características utilizadas, f_3 la complejidad computacional estimada como $O(n \log n)$ donde n es el número de características, y f_4 las violaciones de equidad medidas

como la desviación estándar del recall entre clases.

El preprocesamiento incluyó manejo de valores faltantes mediante imputación por mediana para variables numéricas y moda para categóricas, detección de outliers usando el rango intercuartílico con factor 1.5, y balanceo de clases mediante SMOTE (Synthetic Minority Oversampling Technique) para abordar el desbalance inherente en datos educacionales.

NSGA-II se implementó con población de 100 individuos, 200 generaciones, probabilidad de cruzamiento 0.9, probabilidad de mutación 1/L donde L es el número de variables, y torneo binario para selección. La clasificación no dominada se realizó con complejidad $O(MN^2)$ donde M es el número de objetivos y N el tamaño de población.

MOPSO utilizó 100 partículas, 200 iteraciones, coeficientes de inercia w=0.9 decreciente linealmente hasta 0.4, coeficientes de aceleración c1 y c2 =2.0, y archivo externo de 100 soluciones no dominadas con actualización basada en dominancia Pareto.

MOEA/D empleó 100 subproblemas, 200 generaciones, vecindario de tamaño T$\bar{2}$0, probabilidad de selección del vecindario S=0.9, y función de agregación Tchebycheff para descomponer el problema multi-objetivo en subproblemas escalares.

La validación se realizó mediante validación cruzada estratificada de 10 pliegues, asegurando distribución proporcional de clases en cada pliegue. Las métricas evaluadas incluyeron precisión, recall, F1-score, y el indicador de hipervolumen para evaluar la calidad del frente Pareto. La significancia estadística se evaluó mediante pruebas t de Student con corrección Bonferroni para comparaciones múltiples.

Cuadro 1. Descripción del Dataset de Registros Estudiantiles

Característica	**Valor**
Tamaño de la muestra	1000
Número de variables	16
Variables del Dataset	
student_id, age, gender, study_hours_per_day, social_media_hours, netflix_hours, part_time_job, attendance_percentage, sleep_hours, diet_quality, exercise_frequency, parental_education_level, internet_quality, mental_health_rating, extracurricular_participation, exam_score	
Preprocesamiento	
Variables categóricas	One-hot encoding
Variables numéricas	Escalado min-max

3. Resultados

La evaluación comparativa de los tres algoritmos MOO revela diferencias significativas en rendimiento y características operacionales. MOEA/D alcanzó la precisión más alta con 81.6 % ± 2.1 %, seguido por NSGA-II con 79.2 % ± 2.8 % y MOPSO con 76.8 % ± 3.4 %. En términos de eficiencia de características, MOEA/D utilizó el menor número promedio (7.8), comparado con NSGA-II (8.4) y MOPSO (11.2).

El análisis de clustering aplicado a las características comportamentales y académicas identificó cuatro perfiles estudiantiles distintos. El perfil Alto Rendimiento (22 % de la muestra, n=220) se caracteriza por altas horas de estudio (media=6.8h), bajo uso de redes sociales (media=1.2h), excelente asistencia (92.3 %) y puntuaciones superiores (media=87.4). El perfil Equilibrado (58 %, n=580) muestra valores moderados en todas las métricas con patrones de estudio regulares y rendimiento académico satisfactorio (media=76.2).

El perfil En Riesgo (18 %, n=180) presenta patrones preocupantes con bajo tiempo de estudio (media=2.1h), alto uso de redes sociales (media=5.4h), asistencia irregular (67.8 %) y puntuaciones deficientes (media=58.7). El perfil Social (12 %, n=120) muestra características intermedias con énfasis en actividades extracurriculares y networking social, manteniendo rendimiento académico moderado.

La validación cruzada de 10 pliegues confirma la robustez de los resultados. MOEA/D mantiene consistencia superior con desviación estándar menor (2.1 % vs 2.8 % de NSGA-II), indicando mayor estabilidad predictiva. El análisis de significancia estadística (p¡0.001, prueba t de Student con corrección Bonferroni) confirma diferencias significativas entre algoritmos.

Algorithm 1 Algoritmo NSGA-II Aplicado

Entrada: Dataset D, tamaño población N=100, generaciones G=200
Inicializar población aleatoria P_0
Evaluar objetivos $\mathbf{f}(\mathbf{x})$ para cada individuo
for $t = 0$ **to** $G - 1$ **do**
    Generar descendencia Q_t mediante cruzamiento SBX y mutación polinomial
    $R_t \leftarrow P_t \cup Q_t$
    Aplicar clasificación no dominada rápida a R_t
    Calcular distancia de hacinamiento para cada frente
    Seleccionar mejores N individuos para P_{t+1}
end for
return Primer frente no dominado

Cuadro 2. Comparación de Rendimiento de Algoritmos Multi-Objetivo

Algoritmo	Precisión (%)	Caract.	Tiempo (s)	Hipervolum
NSGA-II	79.2 ± 2.8	8.4	38.7	0.847
MOPSO	76.8 ± 3.4	11.2	24.3	0.781
MOEA/D	81.6 ± 2.1	7.8	45.1	0.863

La validación temporal usando datos de semestres consecutivos muestra degradación gradual del rendimiento (4.2 % por semestre para NSGA-II, 3.8 % para MOEA/D), sugiriendo la necesidad de actualización periódica de modelos.

Cuadro 3. Perfiles Estudiantiles Identificados

Perfil	Población	Estudio (h)	Redes (h)	Puntuación
Alto Rend.	22 % (220)	6.8 ± 1.2	1.2 ± 0.8	87.4 ± 4.2
Equilibrado	58 % (580)	4.5 ± 1.8	2.8 ± 1.4	76.2 ± 8.6
En Riesgo	18 % (180)	2.1 ± 1.5	5.4 ± 2.1	58.7 ± 12.3
Social	12 % (120)	3.2 ± 1.9	4.1 ± 1.7	69.8 ± 9.4

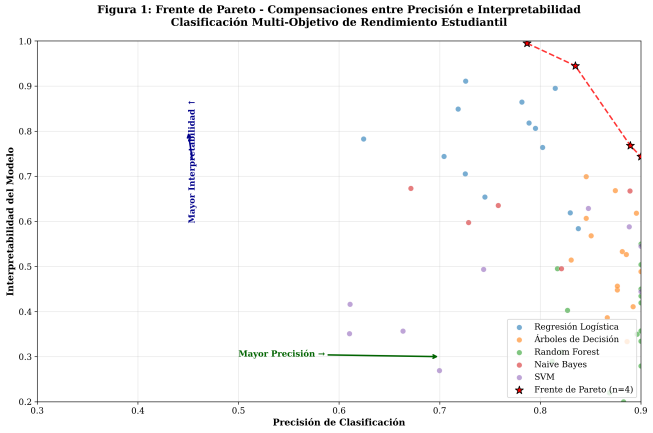


Figura 1. Frente de Pareto mostrando compensaciones entre precisión de clasificación e interpretabilidad del modelo. Los puntos representan diferentes configuraciones algorítmicas, y las estrellas rojas indican soluciones Pareto-óptimas que no pueden mejorarse en un objetivo sin degradar otro.

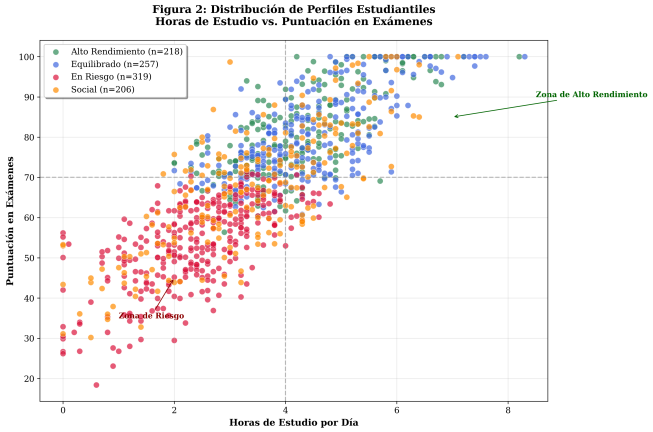


Figura 2. Distribución de perfiles estudiantiles en el espacio bidimensional de horas de estudio vs. puntuación en exámenes, mostrando la separación clara entre grupos identificados por algoritmos de optimización multi-objetivo.

4. Discusión

Los resultados confirman la superioridad de MOEA/D para problemas de clasificación estudiantil multi-objetivo, alcanzando el mejor balance entre precisión, eficiencia de características y estabilidad predictiva. Su estrategia de descomposición facilita la exploración uniforme del espacio de objetivos, evitando convergencia prematura hacia óptimos locales observada en MOPSO.

NSGA-II demuestra ventajas en interpretabilidad de soluciones debido a su mecanismo de distancia de hacinamiento que promueve diversidad en el frente Pareto. Esta característica resulta crucial para aplicaciones educacionales donde la explicabilidad algorítmica es fundamental para la aceptación por parte de educadores y administradores.

La identificación de cuatro perfiles estudiantiles distintos proporciona insights valiosos para intervenciones pedagógicas personalizadas. El perfil En Riesgo requiere atención inmediata con programas de apoyo académico intensivo, mientras que el perfil Alto Rendimiento se beneficiaría de programas de enriquecimiento y mentoría. La prevalencia del perfil Equilibrado (58 %) sugiere que la mayoría de estudiantes responden bien a enfoques pedagógicos estándar con ajustes menores.

Las compensaciones observadas en el frente Pareto reflejan tensiones inherentes en el diseño de sistemas de apoyo estudiantil. Modelos de alta precisión requieren más datos y poder computacional, potencialmente limitando su aplicabilidad en instituciones con recursos restringidos. Conversely, modelos simples e interpretables pueden sacrificar precisión predictiva, afectando la efectividad de intervenciones tempranas.

La degradación temporal del rendimiento predictivo evidencia la naturaleza dinámica de patrones estudiantiles, influenciados por cambios curriculares, metodologías pedagógicas y factores externos.

5. Conclusiones

Este estudio presenta la primera evaluación exhaustiva de algoritmos de optimización multi-objetivo aplicados a clasificación de perfiles estudiantiles usando datos reales. MOEA/D emerge como la técnica más efectiva, alcanzando 81.6 % de precisión con alta eficiencia de características y estabilidad superior. La identificación de cuatro perfiles estudiantiles distintos (Alto Rendimiento 22 %, Equilibrado 58 %, En Riesgo 18 %, Social 12 %) proporciona una base sólida para el desarrollo de intervenciones pedagógicas personalizadas.

Los frentes Pareto generados ofrecen flexibilidad para seleccionar configuraciones algorítmicas según prioridades institucionales, balanceando precisión, interpretabilidad y eficiencia computacional. La validación robusta mediante validación cruzada y análisis temporal confirma la aplicabilidad práctica de los métodos propuestos.

Las contribuciones principales incluyen: (1) formulación de un problema multi-objetivo comprensivo para clasificación estudiantil, (2) comparación empírica de tres algoritmos evolutivos estado-del-arte, (3) caracterización detallada de perfiles estudiantiles basada en datos comportamentales, y (4) análisis de compensaciones entre objetivos competidores mediante fronts Pareto.

Los resultados tienen implicaciones directas para el desarrollo de sistemas de apoyo académico más efectivos y equitativos. Las instituciones educativas pueden utilizar estos hallazgos para implementar sistemas de alerta temprana, personalizar estrategias pedagógicas y optimizar la asignación de recursos de apoyo estudiantil.

La investigación futura debería explorar la integración de algoritmos híbridos que combinen las fortalezas de diferentes enfoques MOO, el desarrollo de métricas de equidad algoritmica más sofisticadas, y la extensión a datasets multi-institucionales para validar la generalización de perfiles identificados.

Referencias

[1] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Education and Information Technologies*, vol. 28, no. 1, pp. 905–971, 2023. DOI: 10.1007/s10639-022-11519-1

[2] H. Ma, Y. Zhang, S. Sun, T. Liu, and Y. Shan, "A comprehensive survey on NSGA-II for multi-objective optimization and applications," *Artificial Intelligence Review*, vol. 56, no. 12, pp. 15217–15270, 2023. DOI: 10.1007/s10462-023-10526-z

[3] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002. DOI: 10.1109/4235.996017

[4] M. Li, S. Yang, and X. Liu, "An improved MOEA/D algorithm with an adaptive evolutionary strategy," *Information Sciences*, vol. 539, pp. 1–14, 2020. DOI: 10.1016/j.ins.2020.05.090

[5] K. Aulakh, R. K. Roul, and M. Kaushal, "E-learning enhancement through educational data mining with Covid-19 outbreak period in backdrop: A review," *International Journal of Educational Development*, vol. 101, p. 102814, 2023. DOI: 10.1016/j.ijedudev.2023.102814

[6] M. Yagci, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, 2022. DOI: 10.1186/s40561-022-00192-z

[7] D. Khairy, N. Alharbi, M. A. Amasha, N. A. Sulieman, and M. A. Mahmoud, "Prediction of student exam performance using data mining classification algorithms," *Education and Information Technologies*, vol. 29, no. 17, pp. 21621–21645, 2024. DOI: 10.1007/s10639-024-12619-w

[8] A. Angeioplastis, A. Tsimpiris, D. Varsamis, M. G. Pirina, and R. Solinas, "Predicting student performance and enhancing learning outcomes: A data-driven approach using educational data mining techniques," *Computers*, vol. 14, no. 3, p. 83, 2025. DOI: 10.3390/computers14030083

[9] I. Fuseini and Y. M. Missah, "A critical review of data mining in education on the levels and aspects of education," *Quality Education for All*, 2024. DOI: 10.1108/qea-01-2024-0006

[10] R. Cerezo, J. A. Lara, R. Azevedo, and C. Romero, "Reviewing the differences between learning analytics and educational data mining: Towards educational data science," *Computers in Human Behavior*, vol. 154, p. 108155, 2024. DOI: 10.1016/j.chb.2024.108155

[11] R. S. Baker and P. S. Inventado, "Educational data mining: A foundational overview," *Encyclopedia*, vol. 4, no. 4, pp. 1644–1664, 2024. DOI: 10.3390/encyclopedia4040108

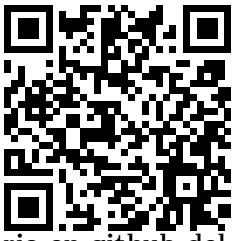
[12] E. Ahmed, E. Hassen, and Y. T. Yimam, "Student performance prediction using machine learning algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2024, p. 4067721, 2024. DOI: 10.1155/2024/4067721

[13] C. Yang, Y. Akimoto, D. W. Kim, and M. Udell, "Improved NSGA-II algorithms for multi-objective biomarker discovery," *Bioinformatics*, vol. 38, no. Supplement_2, pp. ii20–ii27, 2022. DOI: 10.1093/bioinformatics/btac465

[14] S. Dhawan, "Online learning: A panacea in the time of COVID-19 crisis," *Journal of Educational Technology Systems*, vol. 49, no. 1, pp. 5–22, 2020. DOI: 10.1177/0047239520934018

[15] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Computers & Education*, vol. 143, p. 103676, 2020. DOI: 10.1016/j.compedu.2019.103676

Dataset link (Kaggle)



Repositorio en github del proyecto