



Predicción del logro académico en clases espejo: análisis sociodemográfico y pedagógico con minería de datos

Prediction of Academic Achievement in Mirror Classes: Sociodemographic and Pedagogical Analysis through Data Mining

Previsão do desempenho acadêmico em aulas espelho: análise sociodemográfica e pedagógica com mineração de dados

Eilen-Lorena Pérez-Montero¹
Julián-Andrés Quimbayo-Castro²

Recibido: noviembre de 2023

Aceptado: enero de 2024

Para citar este artículo: Pérez-Montero E, L., y Quimbayo-Castro, J. A. (2024). Predicción del logro académico en clases espejo: análisis sociodemográfico y pedagógico con minería de datos. *Revista Científica*, 49(1), 79-98. <https://doi.org/10.14483/23448350.21820>

Resumen

El objetivo de este estudio fue generar modelos predictivos del logro académico en clases espejo en diferentes niveles utilizando el perfil sociodemográfico y pedagógico de 103 estudiantes y algoritmos de minería de datos. Se siguió un enfoque cuantitativo a partir del modelo CRISP-DM (*cross-industry standard process for data mining*), abordando fases como la definición de problema; la adquisición, la comprensión y el análisis de los datos; la extracción de características; el modelado; la evaluación del modelo; y el despliegue. Los resultados muestran que el mejor modelo corresponde al algoritmo *support vector classifier* (SVC), con una precisión del 62 %, una mayor robustez a los cambios y un 0.89 de precisión a la hora de clasificar el logro, lo que representa una reducción de 0.30 con respecto al modelo de regresión logística multinomial (RLOG) y el algoritmo *linear discriminant analysis* (LDA). La métrica *F1 score* presenta un balance de 0.64. Se identificaron atributos predictores importantes para determinar el logro académico respecto a la variable demográfica, como la edad, el género, el semestre y si el estudiante vive con sus padres. Asimismo, en la variable pedagógica se destacaron aspectos como los estímulos a la participación, la puntualidad, el trabajo colaborativo, la claridad de las actividades y su retroalimentación, la facilidad de expresión del docente, los recursos para apoyar el aprendizaje y los medios tecnológicos de comunicación. Este estudio propone herramientas informáticas para que las universidades diseñen estrategias de mejora y prevención respecto al desempeño académico, con base en la implementación de la estrategia de internacionalización denominada *clase espejo* en estudiantes de ingeniería de sistemas de una universidad privada del Huila (Colombia).

Palabras clave: clase espejo; educación superior; logro académico; minería de datos; modelo predictivo.

1. M. Sc. Corporación Universitaria del Huila (Neiva-Huila, Colombia). eilen.perez@corhuila.edu.co.
2. M. Sc. Corporación Universitaria del Huila (Neiva-Huila, Colombia). julian.quimbayo@corhuila.edu.co.

Abstract

The aim of this study was to generate predictive models of academic achievement in mirror classes at different levels using the sociodemographic and pedagogical profile of 103 students and data mining algorithms. A quantitative approach was followed based on the CRISP-DM (cross-industry standard process for data mining) model, addressing phases such as problem definition; data acquisition, understanding, and analysis; feature extraction; modeling; model evaluation; and deployment. The results show that the best model corresponds to the support vector classifier (SVC) algorithm, with an accuracy of 62%, greater robustness to changes, and 0.89 accuracy when classifying achievement, which represents a reduction of 0.30 compared to the multinomial logistic regression model (RLOG) and the linear discriminant analysis (LDA) algorithm. The F1 score metric has a balance of 0.64. Important predictive attributes were identified to determine academic achievement with respect to the demographic variable, such as age, gender, semester, and whether the student lives with their parents. Likewise, in the pedagogical variable, aspects such as stimuli to participation, punctuality, collaborative work, clarity of activities and feedback, ease of expression of the teacher, resources to support learning, and technological means of communication were highlighted. This study proposes computer tools for universities to design improvement and prevention strategies regarding academic performance, based on the implementation of the internationalization strategy called *mirror class* in systems engineering students of a private university in Huila (Colombia).

Keywords: academic achievement; data mining; higher education; mirror class; predictive model.

Resumo

O objetivo deste estudo foi gerar modelos preditivos de desempenho acadêmico em aulas-espelho em diferentes níveis utilizando o perfil sociodemográfico e pedagógico de 103 alunos e algoritmos de mineração de dados. Uma abordagem quantitativa foi seguida com base no modelo CRISP-DM (cross-industry standard process for data mining), abordando fases como definição do problema; aquisição, compreensão e análise dos dados; extração de características; Modelagem; avaliação do modelo; e implantação. Os resultados mostram que o melhor modelo corresponde ao algoritmo classificador de vetores suporte (SVC), com acurácia de 62%, maior robustez às mudanças e acurácia de 0,89 na classificação do alcance, o que representa uma redução de 0,30 em relação ao modelo de regressão logística multinomial (RLOG) e ao algoritmo de Análise Discriminante Linear (ADL). A métrica de pontuação F1 tem um equilíbrio de 0,64. Atributos preditivos importantes foram identificados para determinar o desempenho acadêmico em relação à variável demográfica, como idade, sexo, semestre e se o aluno mora com os pais. Da mesma forma, na variável pedagógica, destacaram-se aspectos como estímulos à participação, pontualidade, trabalho colaborativo, clareza das atividades e feedback, facilidade de expressão do professor, recursos de apoio à aprendizagem e meios tecnológicos de comunicação. Este estudo propõe ferramentas computacionais para universidades desenharem estratégias de melhoria e prevenção do desempenho acadêmico, a partir da implementação da estratégia de internacionalização denominada aula espelho em estudantes de engenharia de sistemas de uma universidade privada de Huíla (Colômbia).

Palavras-chaves: classe espelho; ensino superior; mineração de dados; modelo preditivo; realização acadêmica.

INTRODUCCIÓN

En el contexto académico, en el siglo XXI emerge el fenómeno de la globalización, desencadenando retos y cambios significativos en la ciencia, la tecnología y las comunicaciones. La educación se integra a esta dinámica a través de la internacionalización, buscando formar ciudadanos competentes capaces de promover la integración en ámbitos políticos, económicos, sociales y culturales ([Macazana et al., 2021](#)).

El posicionamiento de las relaciones sociales en la educación superior presenta nuevas modalidades de continua interacción activa. Entre estas están estrategias de internacionalización como las clases espejo. Este enfoque facilita la enseñanza y aprendizaje entre estudiantes con realidades diversas, promoviendo la cooperación e integración global, lo que contribuye a fortalecer la visibilidad, la calidad educativa, la extensión y la investigación en un contexto globalizado ([MEN, 2017](#)).

Implementar una clase espejo implica coordinar dos grupos estudiantiles de distintas ubicaciones, establecer contacto con otra institución educativa, planificar temas comunes, estructurar dinámicas expositivas y horarios, fomentar la participación mediante preguntas y garantizar la conectividad, el *hardware* y el *software* necesarios ([Salas, 2018](#)).

Los estudios de [Yangali et al. \(2021\)](#) señalan que la clase espejo permite fortalecer la competencia investigativa en estudiantes universitarios mediante la conexión con la teoría del aprendizaje colaborativo y la inclusión de las tecnologías de la información y la comunicación (TIC), acercando el mundo exterior desde cualquier lugar de las esferas académicas.

Para [Troncoso \(2022\)](#), una de las dificultades de las clases espejo que se desarrollaron entre la Universidad de Costa Rica (UCR), la Universidad del Valle de Bolivia (UNIVALLE) y la UNITESBA de Celaya, Guanajuato, México, radica en la preparación y la adecuación de los recursos para adquirir conocimiento o alcanzar el logro académico.

[Hederich y Camargo \(2000\)](#) definen el *logro académico* como la síntesis del potencial que un estudiante alcanza durante su exposición al sistema educativo, representada en el conocimiento adquirido a lo largo de su proceso formativo.

La importancia de obtener resultados académicos satisfactorios, la dedicación y la participación en las clases espejo son áreas de interés y estudio en el ámbito universitario, requiriendo la evaluación del agotamiento emocional, la satisfacción temática, las calificaciones y las dinámicas grupales para identificar riesgos ([Orezzoli & Carmen, 2017](#)).

En este contexto, la implementación de la estrategia de internacionalización conocida como clase espejo ha cobrado relevancia, especialmente en los estudiantes de ingeniería de sistemas de una universidad privada en Huila, Colombia. Este estudio no solo busca mejorar su rendimiento académico, sino también enriquecer su experiencia educativa a través de una mirada compartida y una colaboración global.

Existe una gran variedad de factores personales, pedagógicos, sociales e institucionales, que explican los procesos de aprendizaje de un curso formativo ([Franco Hip & Giraldo Ortiz, 2021](#)). Para [Montero y Valverde \(2007\)](#), el factor pedagógico se relaciona con las estrategias utilizadas, los métodos de evaluación, el material didáctico, la comunicación y relación estudiante-docente. El factor sociodemográfico clasifica estudiantes por características diferenciales, vinculando su sexo, nivel económico, tipo de centro educativo y el nivel educativo de padres y su situación laboral.

En la actualidad existen nuevas disciplinas y herramientas, como la minería de datos, que permiten construir modelos analíticos para identificar los patrones descriptivos y las tendencias que caracterizan los comportamientos y logros académicos de los estudiantes ([Hidalgo Cajo, 2018](#)).

[Pérez López y Santín González \(2007\)](#) conceptualizan la minería de datos como como la exploración de relaciones, patrones y tendencias en grandes conjuntos de datos. [Moreno et al. \(1994\)](#) destacan que este campo incluye algoritmos para análisis educativos, aprendizaje automático y predicciones basadas en patrones identificados.

La minería de datos, como lo sugieren [Ayala et al. \(2021\)](#), transforma los modelos de enseñanza-aprendizaje al ofrecer herramientas de análisis, interacción e intervención. Estudios como el de [Rico Páez y Sánchez Guzmán \(2018\)](#) predicen el rendimiento académico de estudiantes de ingeniería mediante un algoritmo basado en calificaciones previas.

Utilizando algoritmos de minería de datos, [Castrillón et al. \(2020\)](#) anticipan el rendimiento académico de estudiantes universitarios, destacando la influencia de la dedicación, la pedagogía docente, la implementación de horarios adecuados, la relación docente-estudiante, la calidad académica y las actividades extracurriculares.

Por su importancia, se han realizado estudios para entender los factores de riesgo relacionados con el logro académico en estudiantes de la UNSAAC, los cuales, mediante algoritmos de aprendizaje automático y datos de ingresos, han identificado factores significativos respecto a características psicosociales, pruebas de ingreso, y datos demográficos, e.g., la escuela profesional, el semestre, el género y la modalidad de estudio ([Candia Oviedo, 2019](#)).

Es así como [Bautista Cañón et al. \(2021\)](#) propusieron un modelo predictivo de aprendizaje supervisado basado en características psicosociales, pruebas de entrada y salida a la educación superior. Utilizando algoritmos como regresión logística, máquinas de vectores de soporte y redes neuronales, los autores alcanzaron una eficiencia cercana al 75 %, evaluando métricas de precisión, *recall* y F1 con porcentajes similares.

Dado el acelerado crecimiento de la tecnología en la educación, este estudio propone un modelo predictivo del logro académico, clasificado en los siguientes niveles: “cumple plenamente”, “cumple en alto grado”, “cumple aceptablemente”, “cumple en bajo grado” y “cumple insatisfactoriamente”. Esto se relaciona con el perfil sociodemográfico y pedagógico de los estudiantes a través de técnicas de aprendizaje automático en las clases espejo de la facultad de ingeniería.

Modelos de minería de datos

La minería de datos es una disciplina crucial en la era de la información que utiliza técnicas estadísticas y algoritmos de aprendizaje automático para descubrir patrones y conocimientos valiosos en conjuntos de datos ([Vieira et al., 2009](#)). En esencia, el aprendizaje automático implica entrenar modelos matemáticos a partir de una base de datos, y su utilidad radica en la capacidad de estos modelos para predecir eventos futuros ([Moreno et al., 1994](#)).

En este sentido, la minería de datos se ha consolidado como una herramienta invaluable en el ámbito académico, brindando la capacidad de extraer perspectivas profundas y significativas de los datos generados por estudiantes, docentes y sistemas de aprendizaje ([Martínez-Abad & Hernández-Ramos, 2018](#)).

La aplicación de esta disciplina en el ámbito educativo ofrece mejoras sustanciales en la toma de decisiones relacionada con la gestión escolar, la identificación de patrones de comportamiento estudiantil, la adaptación de la enseñanza y el aprendizaje a las necesidades individuales, y la predicción del rendimiento académico.

En el marco del estudio que aquí se presenta, se empleó la minería de datos con el propósito de generar modelos predictivos del rendimiento académico en clases espejo en diferentes niveles. Los modelos más influyentes y avanzados en la minería de datos, resaltando sus contribuciones y aplicaciones, son los siguientes:

- **Support Vector Classifier (SVC).** Este algoritmo crea un modelo que asigna los datos a dos o más categorías posibles, representadas por puntos diferentes. El objetivo es encontrar el hiperplano con la máxima distancia (o margen) entre las dos clases, optimizando la clasificación.
- **Linear Discriminant Analysis (LDA).** El LDA se utiliza para reducir la dimensionalidad del espacio de características, manteniendo la información que discrimina entre clases. Este método proyecta los datos de un espacio multidimensional a uno de menor dimensión, con la meta de maximizar la separabilidad entre las clases.
- **Logistic Regression (RLOG).** Este es un método de clasificación que modela la probabilidad de que una determinada entrada pertenezca a una clase. El resultado se representa con una curva en forma de S (función sigmoide) que separa las clases en el espacio de características.

METODOLOGÍA

Esta investigación siguió un enfoque cuantitativo, con un diseño descriptivo que utiliza variables de entrada dadas por el logro académico (variables dependientes) en función de las variables predictoras, *i.e.*, factores sociodemográficos y pedagógicos (variables independientes).

Este estudio se centró en estudiantes de pregrado en ingeniería de sistemas de una universidad privada en Huila, Colombia ($n = 103$), durante el semestre B 2021. La mayoría tenía entre 21 y 25 años (63.1 %). El 21 % tenía 17-20 años, y el 16 % era mayor de 25 años. La muestra incluía 68 hombres y 35 mujeres distribuidos en diferentes semestres, siendo el décimo el más representado (45 %), seguido por el noveno (21 %), el séptimo (18 %) y el quinto (16 %).

Para la recolección de datos, se diseñó un cuestionario con 25 preguntas, adaptado a partir de las investigaciones de [López y Torres \(2018\)](#) y [Montero et al. \(2007\)](#) y enfocado en características psicosociales y factores institucionales, pedagógicos y sociodemográficos asociados al logro académico, el cual se evaluó a través del puntaje final de la clase espejo, registrado en las planillas académicas, con una escala de 0.0 a 5.0, considerando 3.0 como nota aprobatoria. Las especificaciones de las variables se muestran en la [Tabla 1](#).

En lo referente a la minería de datos, se empleó la metodología CRISP-DM (*cross-industry standard process for data mining*), citada en el texto de [Sharma et al. \(2017\)](#), para la construcción de un modelo predictivo ([Figura 1](#)), que cubre 6 fases: 1) comprensión del problema o negocio, 2) adquisición y comprensión de datos, 3) preparación de datos, 4) modelado, 5) evaluación del modelo, 6) despliegue.

RESULTADOS

La implementación del método CRISP-DM (*cross-industry standard process for data mining*) proporciona un marco estructurado para el desarrollo de proyectos de analítica de datos, guiando equipos a través de diversas fases clave, desde la comprensión del negocio hasta la implementación de soluciones. A continuación, se explica cada una de estas fases.

Tabla 1. Especificación de las variables

Variables	Dimensiones	Valores
Sociodemográficas	1. Edad	menores de 17 entre 17 y 20 años entre 21 y 25 años mayores de 25 años
	2. Índice socioeconómico	0 a 7
	3. Género	otro masculino femenino
	4. Semestre	1-10
	5. ¿Vive con padres?	no sí
	6. ¿Activo laboralmente?	no sí
	7. Esfuerzo mental al laborar	no trabaja nunca casi nunca algunas veces casi siempre siempre
	8. Esfuerzo físico al laborar	no trabaja nunca casi nunca algunas veces casi siempre siempre
	9. Número de horas laboradas	no trabaja 1 a 7 h 8 a 16 h 17 a 23 h 24 a 47 h más de 47 h
Pedagógicas	10. Estímulos de participación	nunca casi nunca algunas veces casi siempre siempre
	11. Planteamiento de objetivos	
	12. Preparación de clase	
	13. Necesidades del estudiante	
	14. Puntualidad	
	15. Trabajo colaborativo	
	16. Materiales físicos necesarios	
	17. Claridad de actividades	
	18. Retroalimentación de actividades	
	19. Atención de los docentes	
	20. Profundización temática	
	21. ¿Comparte experiencias con los compañeros?	
	22. Facilidad de expresión del docente	
	23. Ayuda entre compañeros	
	24. Recursos de apoyo	
25. Medios tecnológicos de comunicación		
Logro académico (variable dependiente)	26. Valor cuantitativo de la clase espejo	0.0-5.0

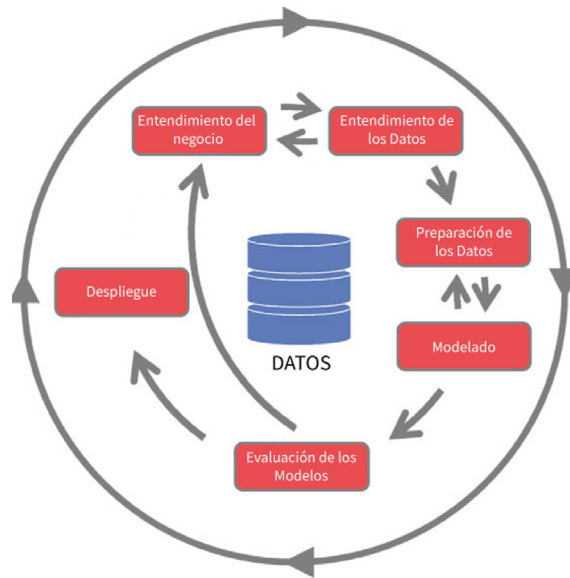


Figura 1. Metodología CRISP-DM

RESULTADOS

La implementación del método CRISP-DM (*cross-industry standard process for data mining*) proporciona un marco estructurado para el desarrollo de proyectos de analítica de datos, guiando equipos a través de diversas fases clave, desde la comprensión del negocio hasta la implementación de soluciones. A continuación, se explica cada una de estas fases.

Desarrollo

Fase 1. Comprensión del problema o negocio. Se evaluó el problema y el contexto académico subyacente. Se procedió a diseñar y validar el cuestionario, analizando la comprensión de los lineamientos para medir los niveles del logro académico en las clases espejo de acuerdo con reglamento de la universidad. Se utilizaron las siguientes categorías: “cumple plenamente” (4.9 a 5.0), “cumple en alto grado” (4.2 a 4.8), “cumple aceptablemente” (3.0 a 4.1), “cumple en bajo grado” (2.0 a 2.9) y “cumple insatisfactoriamente” (0.0 a 1.9).

Fase 2. Adquisición y comprensión de datos. Este proceso inició con la recolección de datos mediante un cuestionario de Google Forms, generando 2678 datos disponibles para el análisis. La fiabilidad del instrumento, evaluada mediante el alfa de Cronbach en Python con un índice de confianza del 95 %, reveló un valor de 0.88. Seguidamente, se evaluó la estructura de los datos, su tipo, formato y significado. La variable que corresponde al logro académico de los participantes se categorizó en tres niveles, "cumple en alto grado", "cumple aceptablemente" y "cumple en bajo grado", considerando que los resultados de la recopilación de información mostraron tres niveles de los cinco definidos en la escala de valoración reglamentada por la Institución educativa.

Debido a esta distribución desbalanceada, se aplicaron técnicas de *oversampling* (aumento de datos minoritarios) y *undersampling* (reducción de datos mayoritarios). Esto aseguró que los algoritmos operaran eficientemente, sin sesgar su desempeño hacia la distribución original.

Utilizando Microsoft Excel, se eliminaron registros con campos nulos en un estudio de variables. La base de datos se convirtió al formato .csv para facilitar su manipulación con herramientas tecnológicas. Se procedió a construir un entendimiento de la tipología de los datos recolectados, dado que la gran mayoría eran de tipo *object* (i.e., texto). Esto se hizo mediante transformaciones, usando el *one hot encoder* o *label encoder* para convertir los datos en valores numéricos o múltiples columnas. Se binarizaron las salidas, como en el caso del género, asignando valores de 1 para una opción y 0 para la otra. No se hallaron datos atípicos o inconsistentes en este ejercicio.

Con el fin de evaluar la normalidad de los datos, se aplicó la prueba de Shapiro-Wilk. Los niveles de significancia de las variables oscilaron entre 0.0000021 y 0.0958000, siendo menores a 0.005 ($p < 0.005$), lo que indica el rechazo de la hipótesis nula.

El análisis exploratorio de la variable sociodemográfica reveló que el 63 % de los sujetos se ubicaban en el índice socioeconómico 2, el 21 % en el estrato 1 y el 16 % restante en los estratos 3, 4 y 5. Se observó que 37 estudiantes hombres no vivían con sus padres, mientras que 30 sí lo hacían. En el caso de las mujeres, 17 no vivían con sus padres y 19 sí.

Respecto a la asociación entre el índice socioeconómico y la condición de vivir con los padres, la mayoría de los estudiantes, específicamente 54, pertenecían a los estratos socioeconómicos 1 y 2. En cuanto a la situación laboral, 68 estudiantes de los estratos 1, 2, 3, 4 y 5 no estaban empleados, siendo los estratos 1 y 2 los más prevalentes en esta categoría.

La [Tabla 2](#) aborda las frecuencias de las dimensiones de la variable pedagógica, i.e., estímulos a la participación, planteamiento de objetivos, preparación de clase, necesidades de los estudiantes, puntualidad, trabajos colaborativos, materiales físicos necesarios, claridad de las actividades y su retroalimentación, atención de los docentes, profundización de temáticas, compartir experiencias con los compañeros, facilidad de expresión del docente, ayuda entre compañeros, recursos de apoyo y medios tecnológicos de comunicación.

La evaluación de los datos reveló patrones consistentes en las respuestas de los participantes, destacando que la mayoría de las frecuencias se situaron en las categorías "casi siempre" y "siempre". En cuanto a los estímulos a la participación, el 80 % de los estudiantes ($n=83$) indicó que los docentes utilizaron efectivamente un sistema de puntos o créditos, sugiriendo una práctica común que beneficia la participación estudiantil y potencialmente mejora el proceso de enseñanza-aprendizaje.

Adicionalmente, para el 82 % de los encuestados ($n=85$) destacó la presencia de un planteamiento de objetivos pedagógicos, evidenciando el compromiso de los docentes con establecer metas educativas claras en el contexto de la clase espejo.

La gran mayoría de los participantes (82 %, $n=85$) mencionaron que los docentes preparaban la clase espejo, lo que es esencial para garantizar la calidad y la coherencia en la presentación de la información y podría beneficiar la experiencia de aprendizaje de los estudiantes.

En relación con la atención de los docentes, un 72 % de los participantes ($n=75$) mostró un fuerte interés y reconoció la importancia de abordar estas necesidades. La puntualidad para comenzar las sesiones, reportada en un 72 %, destaca la importancia de la organización y la gestión eficiente del tiempo en el ámbito académico.

Las estrategias pedagógicas, como el trabajo colaborativo, el material físico, la retroalimentación de actividades, la atención de los docentes, la profundización de temáticas y el uso de medios tecnológicos de comunicación, reportaron valores satisfactorios: 82 % ($n=85$). Esto sugiere un alto grado de eficacia y compromiso en la implementación de estas estrategias, de acuerdo con el proyecto educativo de la institución.

Por otro lado, compartir experiencias con los compañeros y brindarles ayuda reportó una presencia del 76 % en las categorías "siempre" y "casi siempre". Este resultado señala la existencia de áreas de mejora y la necesidad de investigar más a fondo los factores que afectan la participación en este tipo de intercambio no académico.

Fase 3. Preparación de datos. Este proceso correspondió a la construcción del conjunto final de datos, para lo cual se llevó a cabo una rigurosa limpieza y transformación de los datos brutos, identificando y tratando valores atípicos o faltantes. Para adaptar las variables al modelo, se convirtió la variable del logro académico de texto a número, facilitando la implementación de técnicas de balanceo.

Tabla 2. Frecuencias de las dimensiones de la variable pedagógica

Variable pedagógica		Nunca	Casi nunca	Algunas veces	Casi siempre	Siempre
Estímulos a la participación	Masculino	3	2	7	26	30
	Femenino	1	2	5	10	17
Planteamiento de objetivos	Masculino	1	1	10	29	27
	Femenino	1	1	4	9	20
Preparación de clase	Masculino	1	2	10	28	27
	Femenino	1	1	3	10	20
Necesidades de los estudiantes	Masculino	2	6	12	32	16
	Femenino	1	2	5	14	13
Puntualidad	Masculino	2	7	12	31	16
	Femenino	1	2	4	15	13
Trabajos colaborativos	Masculino	2	1	10	29	26
	Femenino	1	1	3	8	22
Materiales físicos necesarios	Masculino	2	2	10	28	26
	Femenino	1	1	2	9	22
Claridad las actividades	Masculino	2	1	12	28	25
	Femenino	1	1	4	8	21
Retroalimentación actividades	Masculino	1	3	8	28	28
	Femenino	2	2	2	9	20
Atención de los docentes	Masculino	2	2	7	27	30
	Femenino	1	2	3	6	23
Profundización de las temáticas	Masculino	2	2	9	27	28
	Femenino	1	1	3	10	20
Compartir experiencias con los compañeros	Masculino	2	3	10	33	20
	Femenino	1	2	6	6	20
Facilidad de expresión del docente	Masculino	2	1	14	32	19
	Femenino	1	1	1	11	21
Ayuda entre compañeros	Masculino	0	2	15	31	20
	Femenino	1	2	4	9	19
Recursos de apoyo	Masculino	0	2	12	27	27
	Femenino	1	2	4	7	21
Medios de comunicación tecnológicos	Masculino	0	2	10	29	27
	Femenino	1	1	4	7	22

Se implementó la técnica de *label encoder*, asignando valores numéricos (0, 1, 2) a las categorías de una columna. El valor 0 indicaba cumplimiento aceptable, 1 denotaba un alto grado de cumplimiento y 2 representa un bajo grado. Se utilizó la librería *scikit-learn*, módulo *preprocessing label encoder*, versión 1.1.2, con Python 3.10 (Figura 2).

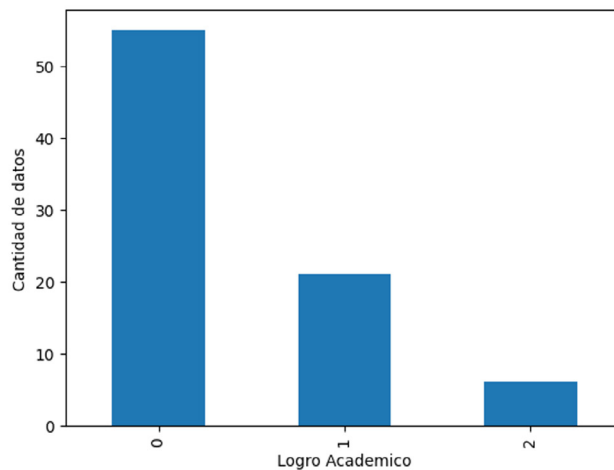


Figura 2. Distribución de la variable logro académico en clases espejo.

Se observó una distribución desigual entre las clases, siendo el tipo 0 predominantemente representado y el tipo 2 menos frecuente. Es crucial considerar esta disparidad para garantizar una representación equitativa al entrenar los algoritmos.

Después de convertir los datos, se aplicó la técnica de validación cruzada con la metodología *Train-Test Split*, en aras de mejorar el modelo predictivo. Se seleccionó una proporción de 80 % para entrenamiento y 20 % para prueba, con un *random state* de 100. Los datos dispersos se normalizaron mediante el procedimiento de mínimos y máximos, reduciendo el ruido y optimizando las variables para la clasificación. Dada la distribución desbalanceada, se emplearon métodos de *oversampling* y *undersampling* para equilibrar las clases, asegurando una clasificación precisa en todo el conjunto de datos. Este enfoque garantiza una implementación efectiva y comprensible del modelado predictivo.

Se empleó la técnica *DecisionTreeClassifier* para obtener la métrica *receiver operating characteristic* (ROC), facilitando la comparación de sensibilidad y especificidad en modelos de clasificación binarios o multclasificados. Los valores de ROC, variando de 0 a 1, deben aproximarse a 1 para obtener un rendimiento óptimo. Se consideraron también la sensibilidad, una métrica que evalúa la capacidad del algoritmo para clasificar correctamente, y la especificidad, que sirve para clasificar correctamente lo negativo. Nuestro enfoque se centró en analizar la clasificación de las clases 0, 1 y 2, representando logros académicos aceptable, alto y bajo, respectivamente.

La métrica ROC proporcionó una visión clara del rendimiento del algoritmo, facilitando la selección óptima. La Figura 3 muestra el área bajo la curva para cada clase; cuanto mayor es el área, mejor es el rendimiento.

En el caso de la clase 0, el equilibrio entre sensibilidad y especificidad fue de 0.65, clasificando correctamente 67 de 103 estudiantes en la categoría de logro académico aceptable. Para la clase uno 1, la curva ROC reveló 0.80, lo que implica que 82 de 103 estudiantes pudieron ser correctamente relacionados

con un logro académico alto. En la clase 2, la métrica indicó 0.42, clasificando correctamente 43 de 103 estudiantes con un logro de bajo grado.

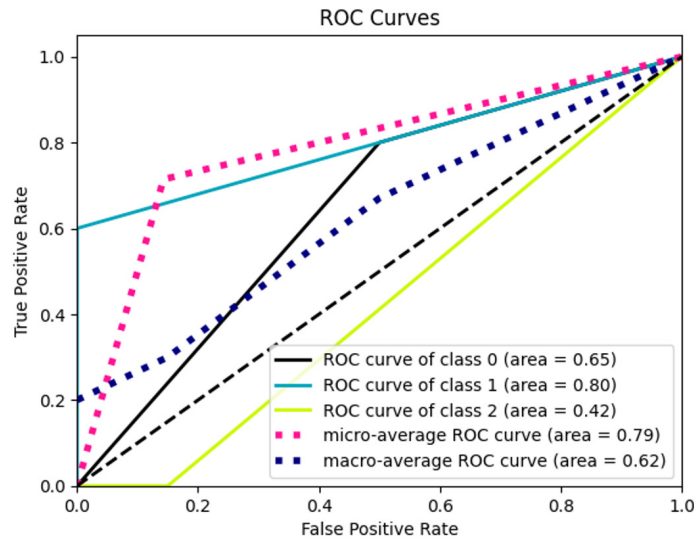


Figura 3. Curva ROC multiclase, logro académico

Si se analiza la curva de precisión y exhaustividad (Figura 4), el modelo desbalanceado destaca al clasificar eficientemente la clase 0, logrando un 78.3 %. Esto implica que 3.1 de cada 4 estudiantes fueron clasificados correctamente en este nivel. La precisión evalúa la calidad del modelo, mientras que la exhaustividad indica la cantidad de personas clasificadas según su clase.

En relación con la clase 1, el modelo logró una precisión del 69.5 %, identificando correctamente 2.8 de cada 4 estudiantes con alto rendimiento. Sin embargo, en la clase 2, solo alcanzó un 4.8 % debido al desbalance de datos, lo que dificultó la detección de bajo rendimiento.

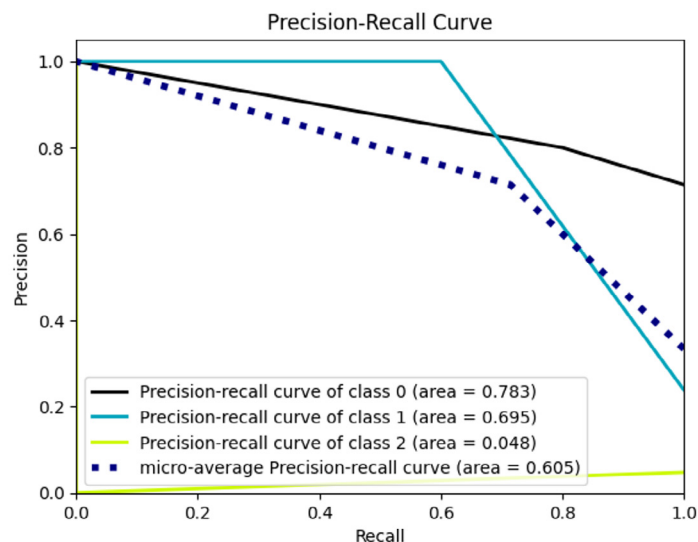


Figura 4. Precisión y exhaustividad por clases, logro académico

Las Figuras 5 y 6 muestran una mejora significativa y generalizada en la sensibilidad y especificidad después de aplicar el balanceo de clases, lo que indica una mayor capacidad del modelo para identificar correctamente los verdaderos positivos y negativos en todas las clases. Aunque la precisión y exhaustividad de la clase 1 disminuyeron en comparación con las demás clases, este resultado puede deberse a una mejor distribución de los datos y a una mayor capacidad del modelo para reconocer patrones en las clases antes desbalanceadas.

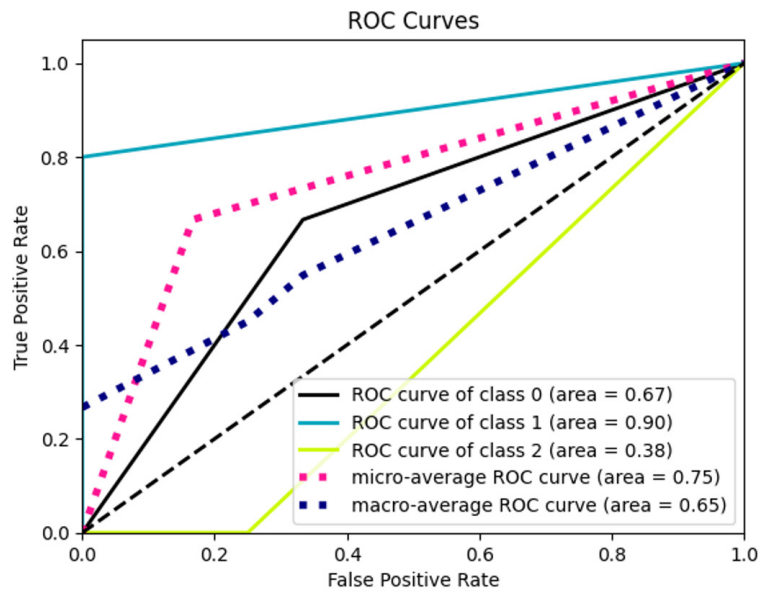


Figura 5. Curva ROC multiclase, logro académico después de aplicar balanceo de clases

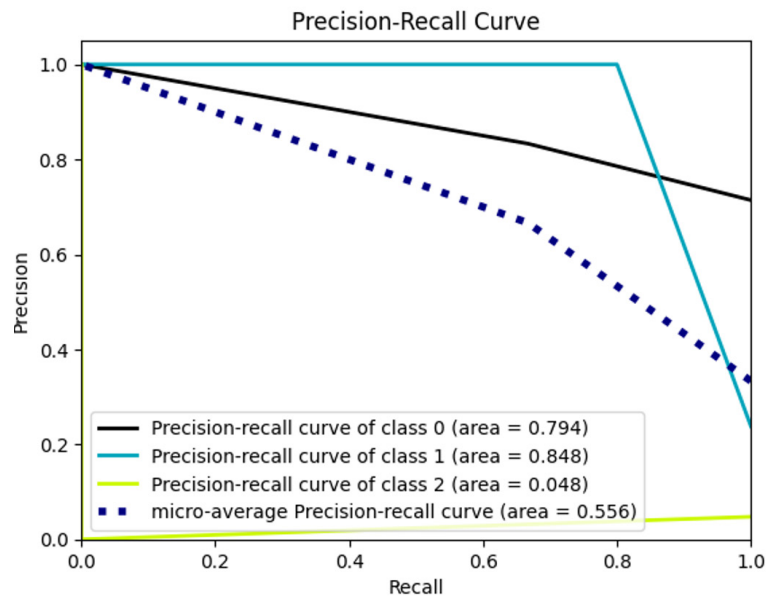


Figura 6. Precisión y exhaustividad por clases, logro académico después de aplicar balanceo de clases

Con el propósito de establecer similitudes y diferencias entre las variables, se empleó la matriz de correlación de las variables predictoras y el nivel del logro académico en la clase espejo. Los valores de correlación proporcionaron información sobre la fuerza y dirección de estas asociaciones, identificando las variables más influyentes. Este análisis permitió determinar qué variables son más significativas, lo que ayudará a ajustar las estrategias pedagógicas para futuras sesiones de formación.

La matriz de correlación de Spearman (Tabla 3) de la variable sociodemográfica, con sus respectivas dimensiones, presenta una correlación positiva moderada entre semestre y edad ($\rho=0.53$), así como entre la variable “¿vive con padres?” y la puntualidad ($\rho=0.53$), y entre “¿vive con padres?” y la ayuda entre compañeros ($\rho=0.55$). Se hace evidente la necesidad de acompañamiento por parte de los padres en la labor pedagógica, pues todas las preguntas de esta dimensión mostraron un rango de 0.27-0.54. Además, se pudo estimar un mayor desempeño en aquellas personas que convivían con sus padres.

El estudio encontró una correlación positiva y robusta con la situación laboral. Se destacan las siguientes correlaciones: labor con esfuerzo mental ($\rho=0.93$), activo laboralmente con esfuerzo físico ($\rho=0.93$), activo laboralmente con número de horas ($\rho=0.95$) y esfuerzo mental con número de horas ($\rho=0.97$). Además, se observó que el hecho de trabajar y el esfuerzo mental y físico muestran las correlaciones más fuertes, que oscilan entre 0.89 y 0.97. Como dirección para futuras investigaciones, se propone analizar cómo las dimensiones relacionadas con el esfuerzo mental y físico en el trabajo pueden influir en el desarrollo académico, particularmente considerando que el horario de estudio suele ubicarse al final de la jornada laboral.

Tabla 3. Matriz de correlación Spearman para las variables sociodemográfica, pedagógica y logro académico

	4	14	23	26
1	0.53	-0.47	-0.33	0.3
5	0.07	0.53	0.55	0.54
26	0.27	-0.27	0.08	1

La Tabla 4 muestra la matriz de correlaciones entre las dimensiones de la variable pedagógica. La mayoría de los casos muestran asociaciones positivas fuertes ($\rho=0.60-0.79$) y positivas muy fuertes ($\rho=0.80-1$).

Tal es el caso de las relaciones positivas muy fuertes entre diferentes aspectos de la interacción en el aula virtual. Por ejemplo, se destaca una correlación notable entre el planteamiento de objetivos y diversas variables, como las necesidades de los estudiantes ($\rho=0.82$) y la puntualidad ($\rho=0.81$). Asimismo, se encuentran relaciones significativas entre la puntualidad y varios aspectos, como la claridad de las actividades ($\rho=0.83$) y la atención de los docentes ($\rho=0.86$). Estas asociaciones sugieren una interconexión compleja entre varios aspectos del entorno educativo, lo que subraya la importancia de abordar múltiples variables para comprender completamente la dinámica de las clases espejo.

Fase 4. Modelado. Se aplicaron técnicas de modelado, ajustando parámetros óptimos a los datos. Se establecieron conjuntos de entrenamiento y prueba (80 y 20 % respectivamente) para evaluar los modelos y verificar la similitud entre los estimadores estadísticos, asegurando la fiabilidad de los resultados. Se utilizó la variable predictora *LogroAca* (logro académico) para poder clasificar la clase de cada estudiante (alto, aceptable o bajo grado).

Tabla 4. Matriz de correlación de Spearman para las variables pedagógica y logro académico

	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
10	1	0.63	0.53	0.53	0.62	0.53	0.57	0.62	0.60	0.68	0.61	0.53	0.59	0.38	0.58	0.56	0.06
11	0.63	1	0.68	0.82	0.81	0.72	0.71	0.72	0.76	0.76	0.73	0.63	0.72	0.68	0.81	0.67	-0.1
12	0.53	0.68	1	0.76	0.64	0.72	0.79	0.67	0.58	0.67	0.66	0.77	0.77	0.69	0.63	0.71	-0.1
13	0.53	0.82	0.76	1	0.81	0.74	0.81	0.79	0.71	0.71	0.69	0.73	0.80	0.72	0.79	0.66	-0.2
14	0.62	0.81	0.64	0.81	1	0.67	0.72	0.83	0.80	0.86	0.79	0.69	0.85	0.67	0.83	0.73	-0.2
15	0.53	0.72	0.72	0.74	0.67	1	0.87	0.72	0.73	0.72	0.76	0.82	0.78	0.73	0.76	0.67	0.01
16	0.57	0.71	0.79	0.81	0.72	0.87	1	0.81	0.75	0.80	0.82	0.56	0.84	0.7	0.73	0.7	-0.01
17	0.62	0.72	0.67	0.79	0.83	0.72	0.81	1	0.9	0.86	0.85	0.70	0.89	0.72	0.81	0.75	0.06
18	0.62	0.76	0.58	0.71	0.80	0.73	0.75	0.9	1	0.88	0.85	0.70	0.85	0.77	0.9	0.82	0.06
19	0.68	0.76	0.67	0.71	0.86	0.72	0.8	0.86	0.88	1	0.86	0.75	0.9	0.68	0.84	0.81	0.08
20	0.61	0.73	0.66	0.69	0.79	0.76	0.82	0.85	0.85	0.86	1	0.79	0.84	0.7	0.79	0.78	0.01
21	0.47	0.63	0.70	0.73	0.69	0.82	0.56	0.70	0.70	0.75	0.79	1	0.70	0.77	0.71	0.75	0.05
22	0.59	0.72	0.77	0.80	0.85	0.78	0.84	0.89	0.85	0.9	0.84	0.72	1	0.70	0.81	0.83	0.01
23	0.38	0.68	0.69	0.72	0.67	0.73	0.7	0.72	0.77	0.68	0.7	0.77	0.70	1	0.60	0.60	0.08
24	0.58	0.81	0.63	0.79	0.83	0.76	0.73	0.81	0.9	0.84	0.79	0.71	0.81	0.60	1	0.88	-0.05
25	0.56	0.67	0.71	0.66	0.73	0.67	0.7	0.75	0.82	0.81	0.78	0.75	0.83	0.60	0.88	1	0.08

El primer algoritmo que se modeló se denomina *linear discriminant analysis* (LDA) y optimiza hiperparámetros que permiten que las predicciones sean resistentes al ruido en la fuente de datos. Se utilizó el teorema de Bayes para estimar la probabilidad condicional, esencialmente calculando la relación entre eventos A y B.

$$P(B|A) = P(AB) / P(A)$$

El algoritmo SVM permite encontrar una superficie de decisión óptima basada en datos, representada por un hiperplano en un espacio dimensional igual al número de atributos del conjunto de datos.

Diversas líneas o hiperplanos pueden separar los datos en dos categorías. La [Figura 7](#) delimita 12 observaciones en un plano, con una línea continua para las clases y líneas discontinuas para los márgenes. Los números 7, 9 y 10 están bien clasificados para la clase azul, mientras que los números del 1 al 6 lo están para la clase roja. Aunque el número 8 está fuera del margen de la clase, pertenece a la clase azul. Sin embargo, el número 11 y 12 están mal clasificados, perteneciendo a la clase roja.

El hiperplano óptimo es aquel que se encuentra más alejado de todas las observaciones, evitando sesgos en la clasificación de nuevos datos. Este enfoque es más robusto ante errores de clasificación, pues emplea penalizaciones para errores fuera del margen del hiperplano. Un valor cercano a cero amplía el margen, permitiendo que varias observaciones violen la clasificación correcta.

También se utilizó el algoritmo RLOG multinomial para predecir y clasificar el nivel de logro académico, identificando las variables influyentes. Se utilizó la muestra de entrenamiento para estimar el modelo con base en 2756 observaciones, creando variables ficticias (*dummies*) que representaban las categorías de cada variable, las cuales eran dicotómicas; mostraban valores de 1 si pertenecían a la categoría y 0 si no).

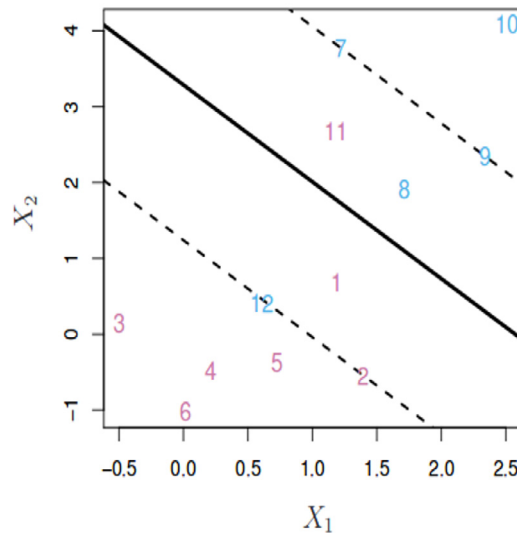


Figura 7. Técnica de modelado SVC

Fase 5. Evaluación del modelo. Se evaluó la calidad de los modelos utilizando métricas estadísticas y siguiendo el enfoque propuesto por [Chapman et al. \(2000\)](#). Los algoritmos entrenados se evaluaron utilizando la matriz de confusión de la librería *Sklearn*, que permitió analizar la precisión de la clasificación binaria (1: desertores, 0: no desertores) mediante verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN).

La precisión en tareas de clasificación evalúa la proporción de TP respecto a la suma de TP y FP. El *recall* indica la capacidad del modelo para identificar TP respecto a la suma de TP y FN. El valor F1 combina precisión y *recall* en una sola métrica, y se calcula como la media armónica entre ambas. La exactitud (*accuracy*) mide el porcentaje de aciertos del modelo, calculada como la proporción de TP y TN respecto a la suma de TP, TN, FP y FN.

Se compararon los resultados de desempeño de los tres modelos supervisados en dos etapas. En la primera etapa ([Tabla 5](#)) se utilizaron todas las variables del cuestionario. El SVC mostró una eficiencia general del 0.62, seguido por el RLOG multinomial con 0.57 y el LDA con 0.43. No hubo diferencias significativas entre los modelos en términos de eficiencia.

En la predicción de la clase 0 (logro académico aceptable), se destacó el algoritmo SVC con una exhaustividad del 0.53, indicando que 2 de cada 4 estudiantes pueden clasificarse en esta categoría según las variables pedagógica y demográfica. El valor F de 0.67 señaló un equilibrio entre precisión y *recall*, lo que significa que 67 de 100 estudiantes pueden ser clasificados correctamente en esta categoría.

Así, el algoritmo inicial mostró un excelente desempeño, pues los datos provenían de una población gaussiana, lo que permitió la separación lineal de las clases. Esto significa que las líneas del hiperplano y sus bordes son totalmente excluyentes y robustos frente a errores de clasificación.

El mejor algoritmo para la case 1 (logro académico de alto grado) fue el LDA, con una exhaustividad del 80 %. Esto implica que, según las variables pedagógica y demográfica, es posible clasificar adecuadamente a 3 de cada 4 estudiantes en esta categoría. Además, el LDA reportó un valor F de 0.67, lo que indica un equilibrio entre precisión y recuperación, permitiendo clasificar correctamente al 67 % de los estudiantes en la categoría 1.

El algoritmo más efectivo para la clase 2 (logro académico de bajo grado) fue nuevamente el SVC, con un equilibrio perfecto de 1.00. Este algoritmo fue el único que logró evaluar tanto precisión como *recall* con un valor de 0.40.

Tabla 5. Métricas iniciales de cada modelo

Algoritmo	Clase original	Clase convertida	Precisión	Exhaustividad	F1	Eficiencia
LDA	Cumple aceptablemente	0	0.80	0.27	0.40	0.43
	Cumple en alto grado	1	0.57	0.80	0.67	
	Cumple en bajo grado	2	0.11	1.00	0.20	
SVC	Cumple aceptablemente	0	0.89	0.53	0.67	0.62
	Cumple en alto grado	1	0.50	0.80	0.62	
	Cumple en bajo grado	2	0.25	1.00	0.40	
RLOG multinomial	Cumple aceptablemente	0	0.80	0.53	0.64	0.57
	Cumple en alto grado	1	0.50	0.80	0.62	
	Cumple en bajo grado	2	0.00	0.00	0.00	

La segunda instancia del análisis ([Tabla 6](#)) involucró el uso de los modelos LDA, SVC y RLOG multinomial con las variables de más alta correlación y los mayores valores de precisión mencionados en la fase anterior.

La eficiencia del LDA aumentó en 0.05, pero la clasificación “cumple en alto grado” desmejoró y perdió equilibrio. La estandarización de datos reveló mejoras en algunos algoritmos, mientras que otros permanecieron sin cambios. Tanto el SVC como el RLOG multinomial mostraron mejoras consistentes en todas las clases y métricas relacionadas con el rendimiento académico en términos pedagógicos y demográficos. El SVC destacó por su alta capacidad de clasificación, mientras que el RLOG multinomial fue más efectivo para clasificaciones de bajo grado.

Fase 6. Despliegue. Este paso integró la utilidad de los modelos en un entorno de operativo. Esto implicó incorporar los modelos y hallazgos en sistemas existentes, realizar pruebas de funcionamiento y garantizar que el despliegue fuera efectivo y sostenible a largo plazo. Además, se consideraron aspectos como la documentación, la capacitación del personal y el monitoreo continuo para asegurar que el sistema funcionara según lo previsto y se logaran los objetivos empresariales.

Para ello, se empleó la biblioteca *Streamlit* de Python 3.10, facilitando la creación de un entorno de despliegue local rápido para visualizar, en tiempo real, el modelo desarrollado en *batch*, *i.e.*, el SVC. Además, se implementó un desarrollo web fácilmente escalable para futura producción en Heroku. Se llevaron a cabo pruebas unitarias simples para evaluar la clasificación de clases según los datos ingresados, las cuales fueron exitosas.

Tabla 6. Métricas de los modelos en la segunda fase del análisis

Algoritmo	Clase original	Clase convertida	Precisión	Exhaustividad	F1	Eficiencia
LDA	Cumple aceptablemente	0	0.83	0.33	0.48	0.48
	Cumple en alto grado	1	0.50	0.80	0.62	
	Cumple en bajo grado	2	0.14	1.00	0.25	
SVC	Cumple aceptablemente	0	1.00	0.47	0.64	0.62
	Cumple en alto grado	1	0.56	1.00	0.71	
	Cumple en bajo grado	2	0.20	1.00	0.33	
RLOG multinomial	Cumple aceptablemente	0	0.80	0.53	0.67	0.62
	Cumple en alto grado	1	0.50	0.80	0.62	
	Cumple en bajo grado	2	0.25	1.00	0.40	

CONCLUSIONES

El modelo propuesto en esta investigación se basó en el UTAUT, centrándose en tres dimensiones esenciales para el logro académico en clases espejo.

Al considerar el acompañamiento de los padres, la preparación docente y los niveles de evaluación (se cumple en alto grado, aceptablemente y en bajo grado), se abordó un amplio espectro de factores sociodemográficos y pedagógicos que influyen en el éxito estudiantil. Estos elementos fueron seleccionados con base en su relevancia y significado para comprender mejor el entorno educativo y su impacto en el rendimiento académico.

El uso de algoritmos de aprendizaje supervisado proporcionó una visión más profunda de las relaciones entre las variables seleccionadas, revelando principalmente una naturaleza lineal en dichas relaciones. Este entendimiento resulta crucial para la aplicación de métodos estadísticos que validen la idoneidad de los datos muestrales. La aplicación de pruebas como la de Shapiro-Wilk permitió confirmar la normalidad de las variables, asegurando así la confiabilidad de los análisis realizados. Esta evidencia fortalece la validez de los resultados y la robustez del enfoque metodológico.

El análisis de correlación de Spearman reveló relaciones significativas entre una amplia gama de variables clave y el logro académico en las clases espejo. Factores como la preparación de clase, el uso de material didáctico, el trabajo colaborativo, el seguimiento del progreso y el uso temáticas actuales demostraron tener un impacto sustancial en el logro académico.

La integración de estos descubrimientos condujo a la creación de tres modelos mediante un proceso *stack*, permitiendo una evaluación exhaustiva que utilizó diversos algoritmos en la misma población. Al

comparar las métricas de estos modelos, se pudo identificar claramente cuál de ellos era más efectivo para predecir el logro académico en este contexto específico.

Estos hallazgos subrayan la importancia de considerar múltiples variables y adoptar enfoques metodológicos complejos para abordar desafíos educativos de manera efectiva. Al comprender mejor cómo estos diversos factores se relacionan entre sí y con el rendimiento estudiantil, podemos mejorar nuestros esfuerzos para apoyar el éxito académico de los estudiantes en las clases espejo y, en última instancia, mejorar la calidad de la educación en general.

El estudio muestra la utilidad de tres algoritmos (LDA, SVC y RLOG multinomial) en el desarrollo de un modelo predictivo. Sin embargo, el SVC se destaca como el más adecuado, con una eficiencia del 62 % y una mayor capacidad para adaptarse a cambios con nuevas observaciones. Esta robustez lo hace consistente en la predicción del logro académico en las tres clases estudiadas.

Además, el RLOG multinomial demostró una eficiencia notable y, junto con la métrica F1, logró un equilibrio óptimo entre exhaustividad y especificidad, garantizando una clasificación precisa, incluso en casos extremadamente positivos o negativos.

Asimismo, se observó que el logro académico tiende a ser más alto entre los estudiantes que viven con sus padres, mientras que trabajar lo afecta negativamente, sugiriendo una posible distracción debido a actividades extracurriculares. En cuanto a la dimensión pedagógica, se determinó que el desarrollo de asignaturas, la metodología del docente y la realización de actividades no representan un riesgo significativo para el logro académico, aunque se sugiere fortalecer estrategias de participación y comunicación entre los actores involucrados.

El desarrollo de clases espejo en estudiantes de ingeniería de sistemas fortaleció el logro académico; el 87 % de los estudiantes obtuvieron una nota aceptable o aprobatoria (3.0 en una escala de 0 a 5.0), lo que indica un proceso formativo positivo. Además, la planificación y ejecución de la clase espejo fueron satisfactorias, pues, según los resultados, se realizaron "siempre" o "casi siempre". Esto va de la mano con lo que sugieren [Franco Hip y Giraldo Ortiz \(2021\)](#) y [Yangali et al. \(2021\)](#).

En general, la propuesta metodológica se aproxima a la teoría de [Escovedo y Koshiyama \(2020\)](#), destacando el uso de técnicas de aprendizaje automático como una herramienta efectiva para desarrollar modelos que clasifiquen el logro académico desde el inicio de las clases espejo. Estos modelos ofrecen apoyo en la toma de decisiones para las instituciones de educación superior.

Esta contribución pretende servir como un punto de partida sólido para futuras investigaciones que exploren en profundidad la aplicación práctica de la minería de datos en la educación superior. Se pretende comprender cómo esta herramienta puede emplearse de manera efectiva para abordar los desafíos específicos que enfrentan las instituciones educativas en la actualidad, con el fin de mejorar los procesos de enseñanza y aprendizaje y fomentar así el éxito académico de los estudiantes.

AGRADECIMIENTOS

Los autores agradecen a la Corporación Universitaria del Huila por su financiamiento y a los estudiantes del programa de Ingeniería de Sistemas por su valiosa colaboración y participación en este estudio.

CONTRIBUCIÓN DE AUTORÍA

Eilen-Lorena Pérez-Montero: investigación, metodología, análisis formal, escritura-borrador original.

Julián-Andrés Quimbayo-Castro: investigación, metodología, supervisión, escritura-revisión y edición.

REFERENCIAS

- Ayala Franco, E., López Martínez, R. E., Menéndez Domínguez, V. H. (2021). Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativos. *Revista de Educación a Distancia (RED)*, 21(66), e463561. <https://doi.org/10.6018/red.463561>
- Bautista Cañón, E., Quirama Salamanca, J. E., Bautista Cañón, E. (2021). Modelo predictivo del progreso en el aprendizaje de los estudiantes de uniminuto aplicando técnicas de machine learning. *Conrado*, 17(83), 305-310.
- Candia Oviedo, D. I. (2019). *Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático*. <http://repositorio.unsaac.edu.pe/handle/20.500.12918/4120>
- Castrillón, O. D., Sarache, W., Ruiz Herrera, S. (2020). Predicción del rendimiento académico por medio de técnicas de inteligencia artificial. *Formación universitaria*, 13(1), 93-102. <https://doi.org/10.4067/S0718-50062020000100093>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. <https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf>
- Escovedo, T., Koshiyama, A. (2020). *Introdução a data science: algoritmos de machine learning e métodos de análise*. Casa do Código.
- Franco Hip, C., Giraldo Ortiz, J. J. (2021). *Guía práctica de planificación y ejecución de clases espejo*. Universidad Peruana de Ciencias Aplicadas. <https://repositorioacademico.upc.edu.pe/handle/10757/657446>
- Franco Hip, C., Giraldo Ortiz, J. J. (2021). *Guía práctica de planificación y ejecución de clases espejo*. Universidad Peruana de Ciencias Aplicadas. <http://hdl.handle.net/10757/657446>
- Hederich, M. C., Camargo, U. Á. (2000). *Estilo cognitivo y logro de aprendizaje en la ciudad de Bogotá*. Centro de investigaciones de la Universidad Pedagógica Nacional.
- Hidalgo Cajo, B. (2018). Data mining in learning management systems in university education. *Campus Virtuales*, 7(2), 115-128.
- Latiesa, M. (1992). *La deserción universitaria: desarrollo de la escolaridad en la enseñanza superior: éxitos y fracasos* (vol. 124). Centro de Investigaciones Sociológicas.
- López Díaz, H., Torres Manrique, D. L. (2018). Características psicosociales de los estudiantes de primer semestre de una universidad privada, facultad de educación. *Investigium IRE Ciencias Sociales y Humanas*, 9(1), 28-39. <https://doi.org/10.15658/investigiumire.180901.03>
- Macazana Fernández, D. M., Romero Díaz, A. D., Vargas Quispe, G., Sito Justiniano, L. M., Salamanca Chura, E. C. (2021). Procedimiento para la gestión de la internacionalización de la educación superior. *Revista Dilemas Contemporáneos: Educación, Política y Valores*, 8, 2585. <https://doi.org/10.46377/dilemas.v8i.2585>
- Martínez-Abad, F., Hernández-Ramos, J. (2018). Técnicas de minería de datos con software libre para la detección de factores asociados al rendimiento. *REXE Revista de Estudios y Experiencias en Educación*, 2(2), 135-145. <https://doi.org/10.21703/rexe.Especial3201812514512>
- Ministerio de Educación Nacional (MEN) (2017). *Internacionalización de la educación superior*. <https://www.mineducacion.gov.co/1759/w3-article-196472.html?noredirect=1>
- Montero Rojas, E., Villalobos Palma, J., Valverde Bermúdez, A. (2007). Factores institucionales, pedagógicos, psicosociales y sociodemográficos asociados al rendimiento académico en la universidad de costa rica: un análisis multinivel. *Electrónica de Investigación y Evaluación Educativa*, 13(2), 215-234.
- Moreno, A., Armengo, E., Béjar, J., Belanche, L., Cortés, U., Gavalda, R., Gimeno, J. M., Martín Muñoz, M., Sánchez, M. (1994). *Aprendizaje automático*. UPC. <https://doi.org/10.5821/ebook-9788483019962>

- Orezzoli , H., Carmen, E. (2017). *Engagement académico en estudiantes de clases espejo en una universidad privada del Perú, durante el año 2016* [Tesis de maestría, Universidad Privada del Norte]. <https://hdl.handle.net/11537/10939>
- Pérez López, C., Santín González, D. (2007). *Minería de datos: técnicas y herramientas*. Thomson.
- Raschka, S., Mirjalili, V. (2019). *Python machine learning*. Marcombo.
- Rico Páez, A., Sánchez Guzmán, D. (2018). Análisis de datos de estudiantes de ingeniería para la predicción del rendimiento académico mediante técnica de clasificación bayesiana. *Dilemas Contemporáneos: Educación, Política y Valores*, 3(3), e548.
- Salas Segura, J. (2018). EMP Clase Espejo el instrumento más práctico para internacionalización e investigación. *Revista Académica Arjé*, 1(2), 26-33.
- Sharma, V., Stranieri, A., Ugon, J., Vamplew, P., Martin, L. (2017). *An agile group aware process beyond CRISP-DM: a hospital data mining case study* [Artículo de conferencia]. International Conference on Compute and Data Analysis, Nueva York, NY, USA. <https://doi.org/10.1145/3093241.3093273>
- Troncoso Colín, P. (2022). Educación digital en instituciones de educación superior (IES): clase espejo. *Management Review*, 7(1), 19-30. <https://doi.org/10.18583/umr.v7i1.195>
- Vieira Braga, L. P., Ortiz Valencia, L. I., Ramírez Carvajal, S. S. (2009). *Introducción a la minería de datos*. E-papers.
- Yangali Vicente, J., Varon Triana, N., Calla Vasquez, K. (2021). Clase espejo como estrategia pedagógica para fortalecer la competencia investigativa en estudiantes de universidades latinoamericanas. *Revista del Instituto de Estudios Superiores en Educación*, 35, 3-21. <https://doi.org/10.14482/zp.35.001.42>

