Analítica de datos y factores de la retención escolar en la educación superior.

Adalberto Mejía-Martinez[1], Harold Valdez-Morales[1], Yaxk'in U Kan Coronado-González[2]

[1] Universidad La Salle, Facultad de Negocios, Ciudad de México, México.
[2] Universidad La Salle, Vicerrectoría de Investigación, Ciudad de México, México.

adalberto.mejia@lasallistas.org.mx, haroldvaldez@lasallistas.org.mx, yaxkin.coronado@lasalle.mx

Resumen. En los últimos años la retención escolar en diferentes niveles del sistema educativo ha disminuido, en particular al nivel superior, esto se traduce en desventajas de la escuela para comprometer a los estudiantes a largo plazo con su educación. Algunos de estos factores se relacionan con la situación socioeconómica, el desempeño académico inicial de los estudiantes, así como situaciones del ámbito familiar y social. Estas problemáticas se presentan de manera conjunta en el estudiante, siendo complejo determinar la retención del alumno de manera individual. Por esto mismo proponemos enforcarnos en el problema de la baja retención escolar en alumnos de nivel universitario, para detectar de manera automatizada la probabilidad de retención de un estudiante, a partir de los factores predominantes, académicos, sociales y económicos que determinan su permanencia en el centro educativo. Se empleo una base de datos de más de 77,000 registros conteniendo evaluaciones, registros socioeconómicos y actividades extracurriculares, se aplica una serie de métodos de aprendizaje automatizado como bosques aleatorios, redes neuronales de clasificación y regresión. Obteniendo una mejor predicción de alumnos no retenidos para los bosques aleatorios de alrededor del 75%. Se puede concluir para el caso de detección de poblaciones subrepresentadas como las no retenidas, los métodos de bosques de decisión presentan un mejor desempeño y explicabilidad de variables al ser totalmente transparentes en su cálculo para grandes bases de datos.

Palabras Clave: Deserción escolar, machine learning, bosques aleatorios.

1 Descripción de la problemática prioritaria abordada

En el sector educativo la gran cantidad de datos generados por las instituciones son de naturaleza muy diversa, como opiniones de alumnos hacia docentes, hasta valores numéricos de las calificaciones de cada periodo, estos datos pueden aportar una visión más global en las causas, motivos y decisiones que llevan a la retención escolar por parte del estudiante, llevando a las autoridades educativas a realizar una mejor toma de decisiones, favoreciendo la retención escolar y la conclusión optima de los estudios del estudiante.

Los datos en el área de la educación son ampliamente valiosos tanto para la detección de la deserción escolar[1] o la nula retención estudiantil, así como casos de aplicación de becas[2], situación socioeconómica de estudiantes o la detección de afecciones emocionales. El reciente uso de técnicas de analítica de datos nos permite explorar, diseñar y ejecutar decisiones más precisas, conociendo la población estudiantil y sus demandas. Ayudando a la generación de acompañamientos personalizados y caracterización de los estudiantes en la institución educativa.

Uno de los aspectos fundamentales para brindar el acceso igualitario y hacerlo sostenible, así como el aumento de las competencias profesionales involucra desde el diagnostico hasta la toma de decisiones por parte de la autoridades educativas, siendo altamente complejo conocer sus implicaciones, actualmente la analítica de datos en conjunto con el aprendizaje automático ofrece una toma de decisiones basada en datos, a partir de datos educativos de los estudiantes que

Memorias del Concurso Lasallista de Investigación, Desarrollo e innovación
Vol. 12, Núm. 2, pp. DHS 167-171, 2025, DOI: 10.26457/mclidi.v12i2.4694 Universidad La Salle México
ADALBERTO MEJÍA MARTINEZ, HAROLD VALDEZ MORALES de INGENIERÍA ECONOMICA Y FINANCIERA, de la FACULTAD DE NEGOCIOS, de la UNIVERSIDAD LA SALLE MÉXICO.
YAXK'IN U KAN CORONADO-GONZÁLEZ fue el asesor de este trabajo.

lograría contribuir a la retención escolar y estrategias para el acceso igualitario a la educación superior ofreciendo una oferta atractiva para una educación sostenible en concordancia con el objetivo del desarrollo sostenible (ODS) 4.3 relacionados al acceso igualitario de todos los hombres y mujeres a una educación técnica, profesional y superior de calidad, incluida la enseñanza universitaria, abonando a la retención escolar para mejorar la calidad educativa y el desarrollo de competencias observadas en el ODS 4.4 enfocado en el aumento del número de jóvenes y adultos que tienen las competencias necesarias, en particular técnicas y profesionales, para acceder al empleo, el trabajo decente y el emprendimiento, contribuyendo a su permanencia en la universidad.

2 Objetivo

Determinar las principales variables, sociales, económicas, familiares o psicológicas que contribuyen la retención escolar en instituciones privadas de educación superior, para la mitigación de la deserción escolar, a través del empleo de algoritmos supervisados de aprendizaje automatizado para la predicción del riego de baja retención escolar.

3 Propuesta teórico-metodológica

Para el proceso de esta investigación se revisó la literatura referente a la analítica de datos en el sector educativo y la implementación de algoritmos de aprendizaje de máquina para la determinación de la retención escolar en nivel superior, la base de datos

Se desarrolló un modelo de analítica de datos para la base de datos del TEC de Monterrey, relacionada a la retención estudiantil entre los años 2015 – 2020, conteniendo variables sociodemográficas, académicas, económicas y familiares[3], siendo la única base digitalizada y curada para la búsqueda integral de variables sociales y académicas para una cantidad enorme de registros anonimizados superando los 70,000 para nivel superior y más de 45 variables Inicialmente se aplicó una exploración de datos estadísticos a las variables, seleccionando las más relevantes para el estudio y un filtrado para el caso de estudiantes universitarios. Posteriormente se aplicaron los modelos de redes neuronales y bosques aleatorios (RF), con árboles de decisión de clasificación y regresión a las variables, tanto en R como Python para la reproducibilidad de resultados independientemente de las condiciones aleatorias del algoritmo, garantizando robustez.

En el caso de redes neuronales se seleccionó un modelo multicapa con una entrada de 64 neuronas, una capa interna de 32 neuronas y una capa de salida de una neurona enfocada en determinar el valor de retención indicado como 1 para un caso positivo y 0 para el caso negativo, este último interpretado como una deserción escolar de la institución.

Para el modelo de bosques aleatorios, se conforma de una serie de árboles de decisiones, los cuales seleccionan la mejor división para clasificar o separar los datos por una regresión lineal, de tal forma que cada característica sea avaluada. Los bosques son la combinación de varios árboles aleatorios, los cuales se selecciona por votación de la mayoría la mejor opción, mejorando la precisión del algoritmo.

Las métricas empleadas para medir y comparar los diferentes modelos aplicados son las siguientes: Exactitud y precisión.

$$E = \frac{TP + T}{TP + TN + FP + FN} \qquad (1)$$

Donde las variables TP, TN, FP y FN corresponden a verdadero-positivo, verdadero-negativo, falso-positivo y falso-negativo en función de la clasificación del problema, en este caso la retención escolar.

Para la precisión se empleó la siguiente definición en función de las clasificaciones correctas.

$$P = \frac{TP}{TP + FP} \qquad (2)$$

Donde las variables son las mismas que en la formula anterior, indicando la clasificación correcta positiva entre la cantidad de valores positivos de la base de datos.

Adicionalmente se calcularon las características más relevantes que influyen en el árbol de regresión, para el entendimiento de las variables más relevantes, como la edad, calificación en pruebas de matemáticas, ingles y otras materias, nivel académico de los padres, genero del alumno, beca académica, etc.

4 Discusión de resultados

En forma de resumen se ejemplifica en la tabla 1 las métricas empleadas y la segmentación de los grupos de retención (1) y no retención (0). Entre los resultados más relevantes está el algoritmo de bosques aleatorios de regresión obteniendo una mayor exactitud y precisión respecto a casos de no retención, en comparación con el uso de redes neuronales. Teniendo la misma precisión para casos de retención.

En este sentido los bosques aleatorios nos brindan una explicación de los factores que influencian la retención escolar de forma individual entre el 7 y 1 por ciento, ver figura 1 por ejemplo en el sector académico donde las pruebas de matemáticas y el examen de ingreso tienen una importancia relevante, posteriormente el promedio del primer semestre es importante para la permanencia y factores económicos como el porcentaje de beca en una institución privada empieza tomar importancia, finalmente los factores sociales y de convivencia académica afectan la retención estudiantil.

Resultados como la habilidad matemática que reflejan los estudiantes de la base de datos, están en acuerdo a trabajos recientes de la actitud hacía las matemáticas al nivel de estudiantes universitarios y la deserción escolar temprana[4]. De igual manera los factores académicos se presentan con mayor relevancia en nivel universitario y factores socioeconómicos o familiares en segunda instancia relevando que factores complementarios a la institución como actividades deportivas y culturales favorecen a una educación experiencia y una retención escolar más allá del desempeño académico[5]. Encontrando que el método de bosques aleatorios es el más adecuado para la detección de deserción escolar o baja retención en nuestro caso[6].

Finalmente es relevante indicar que dada la complejidad de los datos los bosques aleatorios ofrecen, una forma de explicar las relaciones entre la retención estudiantil y los factores subyacentes de la deserción. Siendo por tanto un algoritmo sencillo para la captura de complejidad en datos variados y diversos, llevando a explorar nuevos mecanismos como árboles bayesianos para entender los porcentajes de afectación de cada variable a un estudiante y permitir una predicción cuantitativa y cualitativa, contribuyendo a la mejora de calidad de la educación al identificar patrones claves en la deserción y actuar de manera temprana en casos de riego escolary factores que la provocan a nivel universitario.

5 Conclusiones y perspectivas futuras

El desarrollo de algoritmos para el ámbito educativo se vuelve complejo cuando las variables como la retención se vuelven desbalanceadas al representar arriba del noventa porciento o más de la muestra, representan un reto para seleccionar datos de contraste como la deserción escolar. Para este tipo de problemas la mayoría de los métodos de aprendizaje de máquina no logran capturar toda la población, por lo que se recurre generalmente al balanceo de datos, empleando una selección aleatoria de cada submuestra de población, para asegurar la confiabilidad.

Para nuestro caso de una base de datos educativa, encontramos que el algoritmo de bosques aleatorios logra capturar tanto casos positivos como negativos de la retención escolar, mejorando aún el caso de redes neuronales.

Siendo este un primer acercamiento a la aplicación de métodos de aprendizaje de máquina, los cuales podemos en un futuro mejorar integrando tanto bosques aleatorios como clasificaciones

por clústeres de datos mejorando la precisión y explicar por grupos la retención escolar y sus factores predominantes, aportando a la comprensión y acciones para mejorar la calidad educativa.

4 Agradecimientos

Con extrema gratitud extendemos nuestro reconocimiento y agradecimientos a nuestro asesor y guía, sin el cual nada de esto habría sido posible. El doctor Yaxk´in U Kan Coronado González. También extender nuestro reconocimiento al Centro transdiciplinario de investigación, desarrollo e innovación junto a la vicerrectoría de investigación, por brindarnos la formación, oportunidades y los recursos necesarios para el desarrollo de este proyecto.

5 Referencias

1. Vizcarra V. M. Neuronum, Vol. 10, No. Extra 2, 2024 (Ejemplar dedicado a: Edición Especial -Centro de Investigación Magisterial del Nayar CIMA-MÉXICO-), págs. 259-274.
2. Berlanga V. et. Al Modelo predictivo de persistencia universitaria: Alumnado con beca salario, Educación XX1, Vol 21, 2018, https://doi.org/10.5944/educxx1.20193
3. Alvarado-Uribe, J.; Mejía-Almada, P.; Masetto Herrera, A.L.; Molontay, R.; Hilliger, I.; Hegde, V.; Montemayor Gallegos, J.E.; Ramírez Díaz, R.A.; Ceballos, H.G. Student Dataset from Tecnológico de Monterrey in Mexico to Predict Dropout in Higher Education. Data 2022, 7, 119. https://doi.org/10.3390/data7090119.
4. Geisler, S., Rach, S. & Rolka, K. (2023) The relation between attitudes towards mathematics and dropout from university mathematics—the mediating role of satisfaction and achievement. *Educ Stud Math* **112**, 359–381. https://doi.org/10.1007/s10649-022-10198-6
5. Espinar Álava, Estrella Magdalena, & Vigueras Moreno, José Alberto. (2020). El aprendizaje experiencial y su impacto en la educación actual. *Revista Cubana de Educación Superior*, *39*(3), . Epub 01 de octubre de 2020. Recuperado en 27 de septiembre de 2025, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0257-43142020000300012&lng=es&tlng=es.
6. Rodríguez-Maya, N. E., Lara-Álvarez, C., May-Tzuc, O., & Suárez-Carranza, B. A. (2017). Modeling students' dropout in Mexican universities. *Research in Computing Science*, *139*, 163-175.

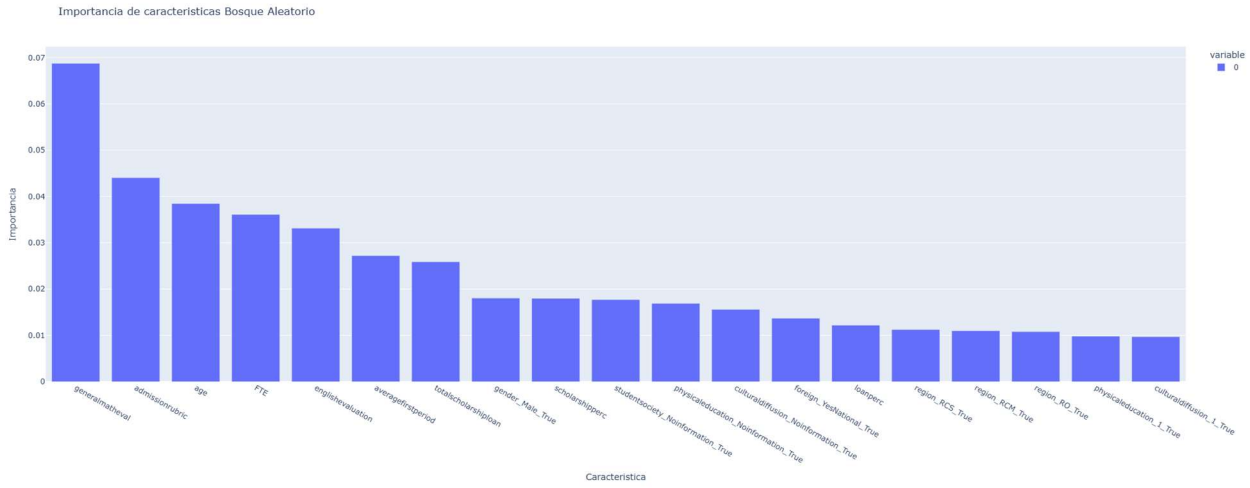

Figura 1. Relevancia de características o variables principales de separación para retención escolar de bosques aleatorios. Elaboración propia.

Tabla 1. Comparativo de modelos implementados para la detección de retención escolar. 0 igual a no retenido, 1 igual a retenido

Modelo	Exactitud (%)	Precisión (0)	Precisión (1)

Bosque aleatorio (clasificación	**26**	**0.25**	**0.27**
Bosque aleatorio (regresión)	**91.78**	**0.76**	**0.92**
Redes neuronales	**88.6**	**0.26**	**0.92**