UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

APLICACIÓN DE TÉCNICAS DE MACHINE LEARNING PARA LA DETECCIÓN DE LAS VARIABLES CAUSANTES DE LA DESERCIÓN ESTUDIANTIL EN LA FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

BOFFILL-BELTRÁN Orestes; orestes.boffillb@uanl.edu.mx

INTRODUCCIÓN

La deserción o abandono estudiantil es un problema reconocido y estudiado mundialmente, que impacta en la efectividad, eficiencia y prestigio de las instituciones educativas, generando consecuencias económicas y/o psicosociales negativas para los estudiantes y sus familias, además de que incide directamente en el desarrollo de un país.

Esta problemática representa una oportunidad para la aplicación de la tecnología y algoritmos de *Machine Learning* a la información académica de los estudiantes.

La detección temprana de los estudiantes en riesgo y la identificación de los principales factores que influyen en la deserción estudiantil han generado una nueva línea de investigación educativa, lo cual permite a las Instituciones de Educación Superior adoptar diferentes medidas para mitigar el fracaso académico.

OBJETIVO GENERAL

Aplicar técnicas de *Machine Learning* para detectar las variables que causan deserción estudiantil en la Facultad de Ciencias Físico Matemáticas de la Universidad Autónoma de Nuevo León

METODOLOGÍA

Figura 1. Flujo de trabajo con objetivos específicos

La información recolectada se estructuró en dos escenarios, tomando como base de datos las calificaciones de los Kardex de estudiantes:

I. Calificaciones hasta terminar el semestre X; X=1,2,3
II. Calificaciones de los Y semestres cursados; Y=1,2,3

TOA01	CalA01	TOA02	CalA02	TOA03	CalA03	TOA04	CalA04	TOA05	CalA05	...	RPT	IRM
1	72	2	72	2	70	2	72	3	80	...	0.732	0.600
1	100	1	88	1	98	1	94	1	70	...	0.900	1.000
3	73	3	97	3	77	3	90	1	85	...	0.844	0.333
1	96	1	100	1	100	1	100	1	96	...	0.984	1.000
1	85	1	90	1	91	1	71	1	92	...	0.858	1.000
2	86	2	73	1	93	1	73	1	70	...	0.790	0.600
1	96	1	100	1	100	1	92	1	100	...	0.976	1.000
1	77	1	74	1	70	3	94	1	79	...	0.788	1.000
1	78	1	70	1	70	1	96	1	75	...	0.778	1.000
1	85	1	85	1	70	1	77	1	100	...	0.834	1.000
1	70	3	70	2	70	1	80	1	74	...	0.728	0.500
3	90	4	70	1	70	2	96	1	79	...	0.810	0.375
1	98	1	87	1	88	1	89	1	78	...	0.880	1.000
1	100	1	89	1	90	1	100	1	94	...	0.946	1.000

Figura 2. Set de datos correspondiente al Modelo Educativo 420 (ME420)

La Figura 2 muestra la base de datos con las calificaciones y variables construidas a través de entrevistas con coordinadores académicos.

Se utilizaron diversos algoritmos de clasificación: *Logistic Regression* (LR), *Random Forest* (RF), *Gradient Boosting* (GB), entre otros.; ajustando sus parámetros mediante herramientas de *Machine Learning*.

Se compararon los algoritmos, y mostramos a continuación resultados de uno de los escenarios.

RESULTADOS

La Figura 3 muestra los valores de la matriz de confusión para LR en uno de los escenarios analizados, con las calificaciones de los estudiantes hasta terminar el primer semestre.

"0": estudiantes que no desertan.
"1": estudiantes que desertan.

Las métricas obtenidas nos indicaron un rendimiento aceptable de LR para la predicción de ambas clases, a pesar del desbalance.

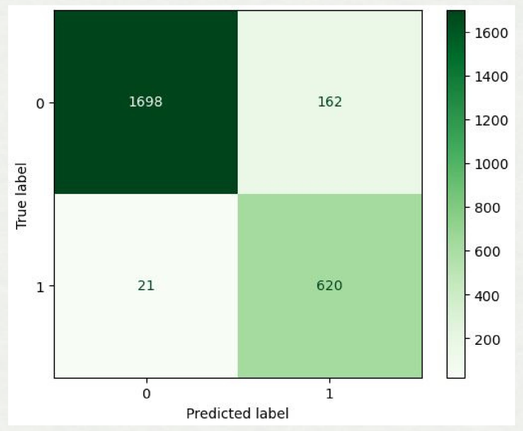

Figura 3. Matriz de confusión para LR

Ambos métodos identificaron las mismas variables significativas en el pronóstico de deserción.

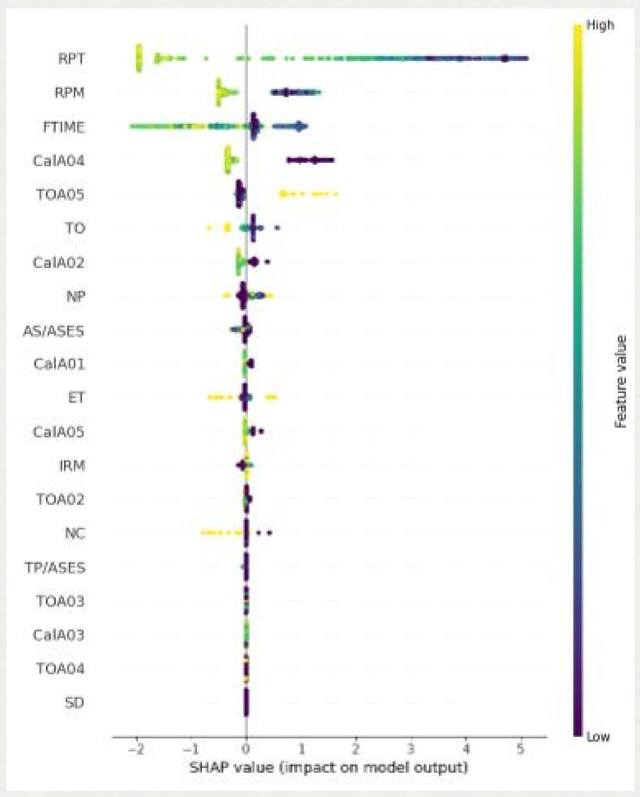

Figura 4. Análisis SHAP para GB

La Figura 4 muestra el análisis SHAP (*Shapley Additive exPlanations*), como resultado de usar GB. Aquí se muestra una medida de la contribución de cada variable a la predicción del modelo.

En este escenario, las variables con mayor influencia corresponden al promedio total, el tiempo en concluir primer semestre, la calificación de Física Básica, entre otras.

CONCLUSIONES

- En particular, los resultados de *Logistic Regression* y *Gradient Boosting*, además de su buen rendimiento, nos facilitaron la interpretación de las variables que afectan a la deserción.
- En el escenario mostrado, identificamos que el promedio en matemáticas, el tiempo que tarda el estudiante en concluir el primer semestre y la calificación de Física Básica, son variables que inciden en la deserción.
- Continuaremos analizando el resto de los escenarios, buscando obtener más información que pueda asociarse a la deserción estudiantil.
- Los resultados anteriores y la literatura, nos indican que si se agregara información referente a las condiciones socioeconómicas de los estudiantes, se podría lograr una mejor interpretación de las causas raíces de la deserción.

REFERENCIAS

1. 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. (2022). Student Dataset from Tecnologico de Monterrey in Mexico to Predict Dropout in Higher Education. Data, 7(9). https://doi.org/10.3390/data7090119
2. Arrieta Matos, L. F. (2021). Diseño de una metodología multicriterio de apoyo a la decisión para la gestión de la permanencia estudiantil de educación superior. Universidad Autónoma de Nuevo León.
3. Palacios-Pacheco, X., Villegas-Ch, W., & Luján-Mora, S. (2019). Application of data mining for the detection of variables that cause university desertion. Communications in Computer and Information Science, 895, 510–520. https://doi.org/10.1007/978-3-030-05532-5_38