PREDICCIÓN DEL ABANDONO EN ESTUDIANTES DE NUEVO INGRESO EN EL CURSO DE MATEMÁTICA GENERAL UTILIZANDO ALGORITMOS DE APRENDIZAJE SUPERVISADO

Autor: José Andrey Zamora Araya

Filiación: Universidad Nacional, Costa Rica

Correo: jzamo@una.ac.cr

Línea Temática: Factores, tipos y perfiles asociados al abandono

Resumen

El objetivo de este trabajo fue determinar, entre tres opciones, cuál es el mejor algoritmo para predecir el abandono en estudiantes de primer ingreso del curso de Matemática General de la Universidad Nacional junto con la identificación de las principales variables asociadas a la predicción. Se utilizaron los algoritmos de Regresión Logística, Random Forest y XGBoost, la métrica de rendimiento elegida fue la F1 Score. El archivo de entrenamiento está constituido por la matrícula de los años 2017 y 2018 y el archivo de prueba por la matrícula del año 2019, los hiperparámetros de los algoritmos se ajustaron por medio de una validación cruzada 5 folds. Una vez ajustados los hiperparámetros se realizó una comparación de las medias de la métrica F1 con un ANOVA de medidas repetidas con validación cruzada 10 folds. Las variables más importantes para la predicción fueron la nota en la PAA, el IDS, la edad y sexo del estudiantado, edad y grado académico de la persona docente, el estrato educativo y la carrera. Los resultados muestran que no existen diferencias significativas en cuanto al rendimiento de los algoritmos (p =0,118) en la métrica seleccionada, por lo que se aconseja utilizar el algoritmo menos complejo que es el de Regresión Logística para interpretar los resultados.

Descriptores o Palabras Clave: Abandono, Modelos Predictivos, Matemática, Aprendizaje Supervisado

Contextualización:

El curso de Matemática General en la Universidad Nacional (UNA) presenta uno de los niveles de aprobación y permanencia más bajos de la institución, pues en promedio, menos de la tercera parte del estudiantado logra aprobar el curso y una cantidad similar no logra concluirlo, es decir, se cataloga como desertor. En cuanto a la definición conceptual del abandono, según Tinto (1982) no existe una definición que pueda abarcar en su totalidad la complejidad del término, por lo que las personas investigadoras lo operacionalizan en función del contexto y los objetivos de investigación. En el caso de la Universidad Nacional, por normativa institucional, se establece que el estudiantado matriculado en un curso que obtiene una nota menor o igual a 2 (en escala de 0 a 10) se considera desertor de la asignatura, pues se asume que en ese rango de calificaciones se encuentran las personas que se desincorporaron del curso, y esta es la definición de abandono que establece para este estudio.

Esta situación provoca rezago académico, lo que a su vez va acompañado de un aumento en la demanda del curso que, dada la situación presupuestaria de la organización, no siempre es posible satisfacer. Asimismo, la desincorporación en el curso es un indicador del abandono temprano a nivel institucional.

Aunque existen programas de apoyo estudiantil como tutorías, la problemática se mantiene. En este contexto, un modelo que permita predecir el abandono en el curso ayudaría a guiar las intervenciones para atender de una manera oportuna al estudiantado con mayor probabilidad de abandono, al identificar aquellas características que hacen que una persona sea más propensa al abandono.

Al respecto, en la actualidad los algoritmos de inteligencia artificial más precisamente el área denominada Machine Learnign (ML), dispone de una serie de técnicas que permiten predecir la ocurrencia de ciertos fenómenos. En el ámbito educativo una subárea de ML llamada aprendizaje supervisado, resulta particularmente útil a la hora de realizar predicciones en contextos relacionados con el abandono y la permanencia.

De acuerdo con Mahesh (2020) los algoritmos de aprendizaje supervisado son aquellos que necesitan de asistencia externa para resolver problemas de regresión o clasificación, en donde se dividen los datos de entrada en dos conjuntos, uno llamado de entrenamiento y otro denominado de prueba, de forma tal que el algoritmo aprende algún tipo de patrón que tienen los datos de entrenamiento y los aplica a los datos de prueba con el objetivo de predecir o clasificar alguna variable de interés, en este caso predecir si un estudiante abandona o permanece en el curso de Matemática General. Par los lectores que deseen profundizar en los métodos de aprendizaje supervisado se recomienda consultar los trabajos de Kuhn y Johnson (2016) y Burkov (2019).

Cómo existen muchos algoritmos en el aprendizaje supervisado, es común realizar los análisis con varios de ellos y luego comparar su poder predictivo y así seleccionar el que presenta mejores resultados. Para ello se elige una métrica de comparación, entre las más comunes están *accuracy* (el porcentaje total de casos clasificados correctamente), precisión (porcentaje de valores clasificados como positivos que son realmente positivos), *recall* (porcentaje o ratio de valores positivos clasificados correctamente) y F1 Score o F1 (Cao, *et al*., 2020). La puntuación F1 se eligió porque es una combinación de precisión y *recall* que logra una comparación más clara y menos sesgada, en particular ante la presencia de datos desequilibrados.

Los estudios al respecto muestran la incidencia que pueden tener las variables académicas, individuales y socioeconómicas a la hora de predecir el abandono escolar (Calva et al., 2021; López- Zambrano et al., 2021, Solís et al., 2018), en particular para los cursos del área de ciencias y matemáticas (Kilian et al., 2020; Muñoz-Camacho, 2018). En el caso de la UNA los estudios se han concentrado en la identificación de factores como el nivel de conocimientos previos, el sexo del estudiantado, la carrea seleccionada y las expectativas sobre ella (Castillo et al, 2020; Zamora-Araya y Villalobos-Madrigal, 2018), más no en el uso de ellos como predictores que puedan contribuir al planteamiento de acciones institucionales tendientes a reducir la magnitud del problema, en particular para el estudiantado proveniente de sectores sociales vulnerables, pues en la institución no se ha profundizado en estudios sobre modelos predictivos aplicables al problema del abandono escolar.

En este particular, la UNA ha implementado un sistema de admisión que favorece el ingreso de estudiantes provenientes de colegios con carencias en cuanto a infraestructura, accesibilidad, recursos económicos y oportunidades educativas. El sistema considera por un lado las características de la de los colegios (modalidad de estudios, horario, tipo de colegio y la ubicación geográfica) y por otro el rendimiento previo en la educación secundaria y en la prueba de actitud académica (PAA) o prueba de admisión (Universidad Nacional, 2022).

El proceso de admisión establece tres estratos, los cuales han sido definidos según las características del colegio de procedencia en:

Estrato 1: colegios privados, subvencionados, bilingües experimentales, científicos, humanistas y extranjeros.
Estrato 2: colegios académicos o técnicos profesionales diurnos y públicos.
Estrato 3: colegios rurales, nocturnos, con bachillerato a distancia, entre otras modalidades.

Durante el proceso de admisión se realiza una ponderación que permite al estudiantado con las notas más altas del estrato 3, el cual representa a las personas provenientes de contextos con menores oportunidades educativas, poder competir por cupos en las diferentes carreras que oferta la universidad con los otros dos estratos.

No obstante, hasta la fecha, no se han implementado acciones en el curso de Matemática General dirigidas específicamente a atender a este grupo estudiantil que, por su contexto educativo previo, presentan deficiencias en cuanto al nivel de conocimiento matemático requerido para un curso introductorio en el área a nivel universitario. Cabe resaltar que, en cuanto a rendimiento en Matemática, los estratos 2 y 3 son los que presentan mayores dificultades y representan alrededor del 80% de la matrícula del curso, cerca de 1 000 estudiantes, sólo durante el primer semestre.

Debido a restricciones presupuestaria, la universidad no puede atender a toda la población que requiere apoyos en el área de Matemática, con tutorías por ejemplo. De esta manera, un modelo capaz de predecir con cierto grado de precisión el abandono en el curso es una herramienta que podría usarse para identificar a estudiantes que requieran atención temprana, al priorizar los apoyos no sólo basados en el estrato, sino en la probabilidad de abandonar el curso y así evitar el abandono en el curso.

En otras palabras, el contar con un algoritmo que pueda predecir la probabilidad de que un estudiante abandone el curso, con el mayor grado de precisión posible, no solo permitirá optimizar la selección de las personas que reciben los apoyos sino además identificar el grado en que otras variables disponibles en los registros universitario contribuyen a predecir el abandono. Por ende, se busca no solo determinar la probabilidad de abandono del estudiantado antes de iniciar el curso, sino también cuáles variables son las más relevantes a la hora de realizar la predicción de si un estudiante abandono o permanece en el curso de Matemática General.

Objetivo

El objetivo del estudio fue determinar cuál es el mejor algoritmo para predecir el abandono en estudiantes de primer ingreso del curso de Matemática General de la Universidad Nacional y, además de calcular el aporte a la predicción del abandono estudiantil de las principales variables asociadas al abandono en el curso, para valorar su uso en futuras intervenciones educativas.

Método:
La investigación es de tipo correlacional de corte longitudinal cuyos participantes son las cohortes estudiantiles de los año 2017, 2018 y 2019 que matricularon el curso de Matemática General para predecir el abandono. La variable dependiente fue el abandono en el curso (clase positiva o categoría a predecir) determinada por el criterio que usa la universidad para operacionalizar o identificar el abandono en un curso, el cuál es la obtención de una nota final de 2.0 o menor (en escala de 0 a 10), y como predictores las variables recolectadas durante el proceso de admisión como la PAA, notas del colegio de procedencia, estrato, entre otras.

Los datos fueron suministrados por el Departamento de Registro y la Escuela de Matemática de la UNA y se excluyeron de los análisis los registros estudiantiles de las personas que realizaron retiro justificado del curso, además de los registros cuyas notas finales de curso fueron reportadas como cero o con la sigla NA, ya que representan al estudiantado que no asistió a clases, pero no realizó formalmente el retiro justificado.

Inicialmente los archivos de datos contenían alrededor de 148 variables, que son las variables que se encuentran disponibles en las bases de datos, tanto del Departamento de Registro como de la Escuela de Matemática y que literatura señala que están potencialmente asociadas al fenómeno del abandono. No obstante, se descartan aquellas variables que brindan información redundante o del todo no resultaban relevantes para el abandono estudiantil. Además, para aquellas las parejas de variables altamente correlacionadas ($r > 0.80$) solo se consideró una de ellas. Al final de este proceso, se seleccionaron un total de 22 variables asociadas con el abandono en el curso de Matemática General. Finalmente, se aplicó un algoritmo envolvente de selección de características llamado Boruta, que genera una medida de importancia para cada una de las variables predictores con respecto a la variable a predecir, en este caso el abandono, de tal forma que se pueden elegir dentro de un gran número de variables sólo aquellas con verdadero potencial para realizar la predicción (Ruto, 2022). Esto redujo la cantidad de variables de 148 a 15, lo que permite mayor interpretabilidad a los resultados.

Para realizar la predicción se evalúo el rendimiento de tres algoritmos de aprendizaje supervisado (Regresión logística, Random Forest y XGBoost) en la métrica F1 Score. Se eligió esta métrica, dado que representa un balance entre precisión y sensibilidad. Además, se utilizan los registros universitarios que contienen variables recolectadas durante el proceso de admisión a la universidad, solicitados por la Dirección de la Escuela de Matemática vía oficio, en el cual se compromete a respetar el anonimato de las personas y cumpliendo los criterios establecidos en la ley 8968 de protección de la persona frente al tratamiento de sus datos personales. Para el conjunto de entrenamiento se tomó la matrícula de los años 2017 y 2018 y como conjunto de entrenamiento la matrícula del año 2019.

Para cada conjunto de datos (entrenamiento y prueba) se implementó un proceso de validación cruzada 5 folds, el cual permitió ajustar los hiperparámetros en los algoritmos de Random Forest y XGBoost para luego ser evaluados en el conjunto de prueba. Recordemos que el conjunto de entrenamiento se usa para que el algoritmo "*aprenda*" el patrón subyacente de la categoría o probabilidad a predecir y con los que se pueden ajustar los hiperparámetros; mientras que el conjunto de prueba se usa para realizar una evaluación del ajuste final del modelo con datos que los algoritmos, hasta el momento, "*no han visto*".

En el caso de la Regresión logística, al no tener que estimar hiperparámetros, solo se trabajó con los conjuntos de entrenamiento y prueba. Para comparar el rendimiento de los algoritmos

se utilizó un ANOVA de medidas repetidas, con la misma semilla aleatoria. Los análisis se realizaron con el software R versión 4.2.1 (R Core Team, 2022), en particular con el paquete mlr3 (Lang et al., 2019).

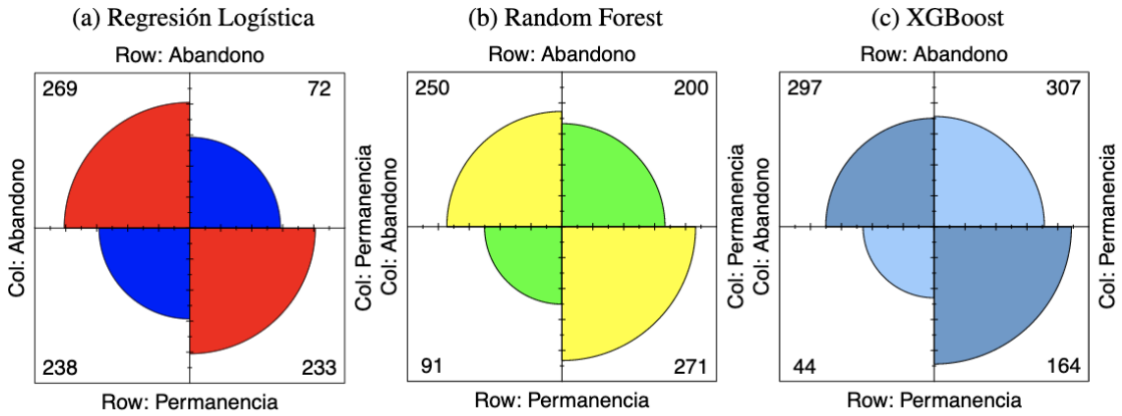
Resultados:

El archivo de entrenamiento estuvo constituido por 1432 estudiantes que matricularon el curso de Matemática General durante los años 2017 y 2018 y el archivo de prueba por 812 estudiantes que matricularon el curso durante el año 2019. Esta división se hace con la intención de representar de la manera más realista posible el escenario en donde eventualmente se aplicaría la predicción, es decir, usando como base la información de años anteriores para predecir el abandono de una nueva cohorte estudiantil.

Una vez ajustados los hiperparámetros en el archivo de entrenamiento se procedió a realizar la evaluación en el archivo de prueba y los resultados en la métrica F1 fueron muy similares en los tres algoritmos: Regresión Logística, Random Forest y XGBoost con valores de 0,6344; 0,6321 y 0,6286 respectivamente. Las matrices de confusión se muestran en la Figura 1.

Figura 1

UNA: Matrices de Confusión según los Algoritmos Utilizados para Predecir el Abandono Estudiantil en el Curso de Matemática General. Validación Año 2019

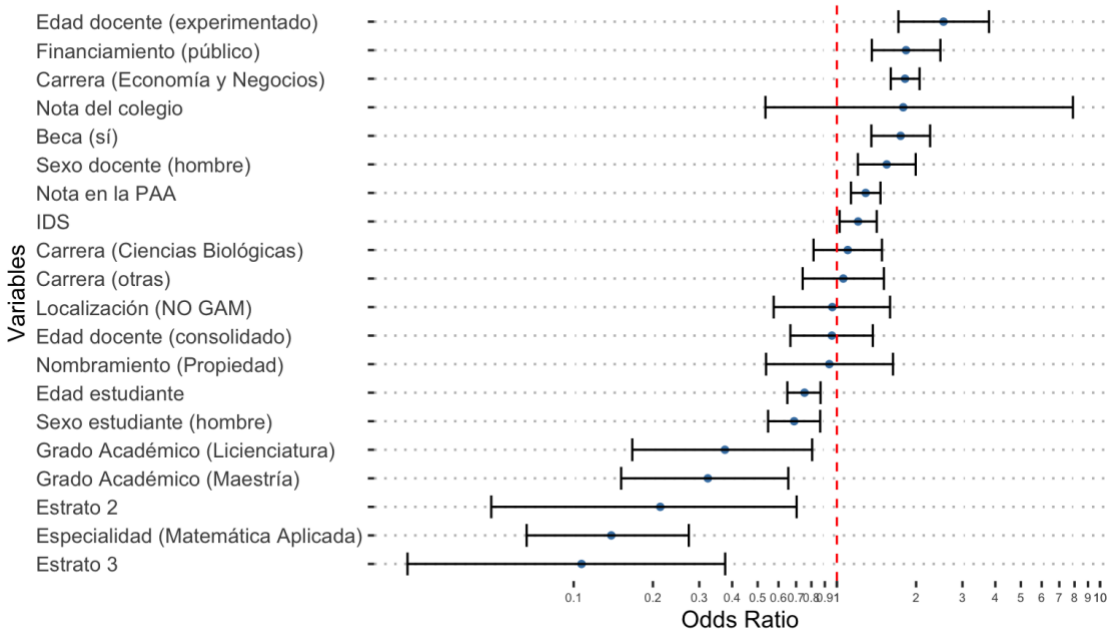


Fuente: Elaboración propia

Una vez que el rendimiento predictivo fue determinado, se procedió a determinar cuáles de las 15 variables son las más relevantes para la predicción, para lo cual se utilizó la medida de importancia de Gini. Para la Regresión Logística y el Random Forest se presentan las medidas de importancia de las 15 variables (en el caso de la Regresión Logística con sus respectivas categorías para las variables cualitativas) y para el XGBoost, el algoritmo solo reportó a 10 variables con un valor significativo para la predicción. Los resultados que muestran las principales variables predictoras del abandono, de acuerdo con cada algoritmo, se aprecian en las figuras 2, 3 y 4.

Figura 2

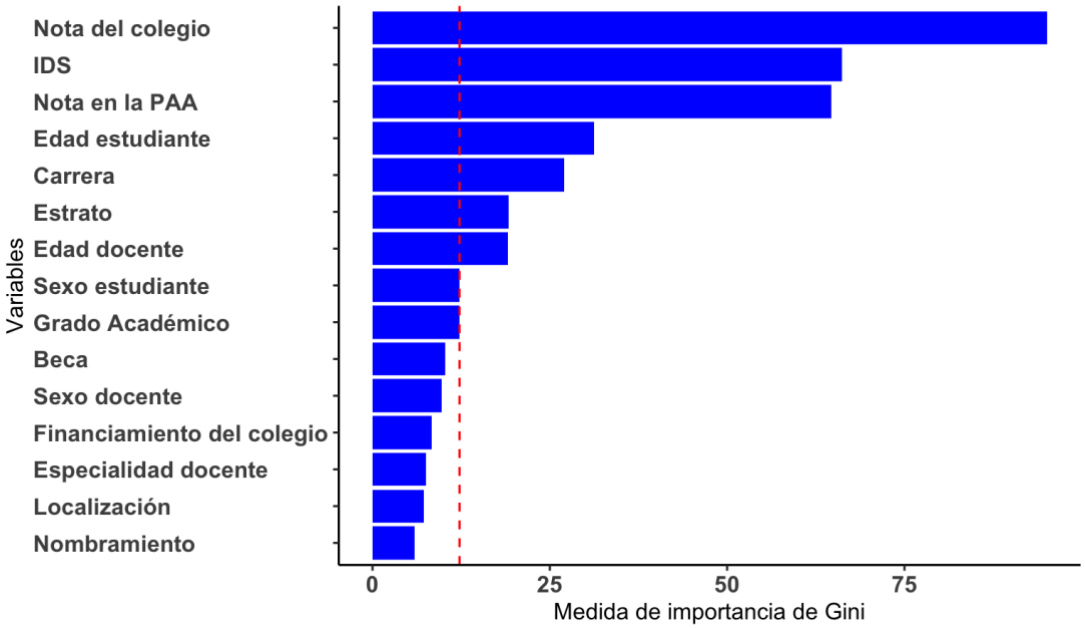
UNA: Importancia de las Variables para el Estudiantado de Primer Ingreso Según el Algoritmo de Regresión Logística. Validación para el año 2019



Fuente: Elaboración propia

Figura 3

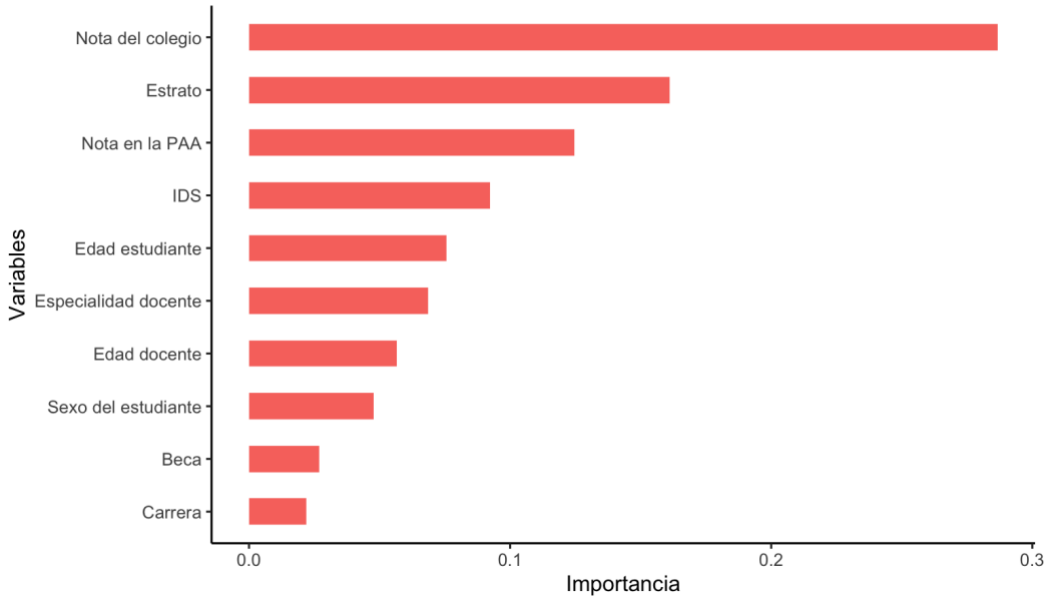
UNA: Importancia de las Variables para el Estudiantado de Primer Ingreso Según el Algoritmo Random Forest. Validación para el año 2019



Nota: La línea punteada representa la mediana de la medida de importancia.
Fuente: Elaboración propia

Figura 4

UNA: Importancia de las Variables para el Estudiantado de Primer Ingreso Según el Algoritmo XGBoost. Validación para el año 2019



Fuente: Elaboración propia

Como puede observarse, las variables más importantes para efectos de predicción fueron la nota del colegio, la nota en la PAA, la edad de la persona docente y el IDS que tienen una relación inversa con el abandono en el curso, es decir, a mayor puntuación menor probabilidad de abandono. Por su parte, entre mayor sea la edad del estudiantado mayor será su probabilidad de abandono.

En lo que respecta a las variables categóricas los hombres son los que tienen mayores probabilidades de abandono, al igual que pertenecer a los estratos 3 y 2 en comparación con el estrato 1. El estudiantado que pertenece a las carreras del área de Economía y negocios es que tiene menor probabilidad de retirarse de las aulas en comparación con el estudiantado de la carrera de administración.

Por su parte el estudiantado que recibe clases con profesores de sexo masculino tiene menor probabilidad de abandono. En contraste el estudiantado que recibe lecciones con el profesorado cuyo grado académico es de licenciatura o maestría tiene mayores probabilidades de abandono en comparación de aquel que recibe clases con el profesorado que tiene el grado de doctorado (ver Figuras 2, 3 y 4).

Adicionalmente, la regresión logística al igual que el XGBoost identificaron como variables relevantes a la beca y la especialidad docente. El estudiantado con beca tiene menor probabilidad de abandono al igual que el que recibe clases con profesores cuyo grado académico está relacionada con el área de la educación, en comparación con aquel que recibe clases con el profesorado cuya área de especialización es la matemática pura o aplicada.

Finalmente se realizó un análisis de varianza con la misma semilla aleatoria utilizando validación cruzada 10-folds, para determinar si existían diferencias significativas en el rendimiento de los algoritmos en la medida F1. Las pruebas de normalidad se realizaron mediante la prueba de Shapiro-Wilk y se verificó el supuesto de normalidad para la medida F1 en los tres algoritmos (Regresión Logística, Random Forest y XGBoost) con valores p de 0,6555; 0,1858 y 0,7816 respectivamente. También se comprobó el supuesto de homocedasticidad de varianzas mediante la prueba de Levene con un valor p de 0,5984.

Tabla 1

UNA: Análisis de Varianza para la Comparación de Medias de los Algoritmos de Regresión Logística, Random Forest y XGBoost en la Predicción del Abandono en el Curso de Matemática General

Fuente de variabilidad	Grados de libertad	Suma de cuadrados	Cuadrado medio	Valor F	Pr(>F)
Algoritmos	2	0,0004	0,0002	0,1180	0,8890
Error	27	0,0514	0,0019		
Total	29	0,0518	0,0021		

Fuente: Elaboración propia

Como se aprecia en la Tabla 1, el rendimiento medio de los tres algoritmos en la medida seleccionada, puntuación F1 Score, no es significativamente diferente (p= 0,8890), por lo que no existe evidencia de que algún algoritmo sea mejor que otro para efectos de predecir el abandono en el curso de Matemática General.

Conclusiones

El abandono estudiantil es un tema complejo y multifactorial, no obstante, la identificación de variables con potencial predictivo por medio de técnicas de aprendizaje supervisado es de gran utilidad, ya que por una parte permiten asignar una probabilidad de que una persona estudiante deserte y con ello, eventualmente pueden ser utilizadas para plantear posibles intervenciones que ayuden a enfrentar esta problemática como se ha hecho en muchas instituciones de educación superior, gracias al uso de herramientas tecnológicas (Contreras et al., 2020; Kuz, 2023; Smith y Gutiérrez, 2022), al usar la probabilidad de abandono como un indicador para focalizar la asistencia.

En el caso del curso de Matemática General de la UNA, se logró identificar a la nota del colegio y la nota en la PAA como variables asociadas al rendimiento previo del estudiantado que son importantes para la predicción del abandono, en concordancia con otros estudios realizados en la institución (Castillo-Sánchez et al., 2020; Hidalgo et al., 2019; Rodríguez-Pineda, 2018). Por otra parte, variables asociadas con el contexto educativo y socioeconómico del estudiantado como el Índice de Desarrollo Social (IDS) y el estrato, también resultaron relevantes; lo que muestra cómo el estudiantado provenientes de sectores socioeconómicos desfavorecidos y con menores oportunidades educativas (estrato 3) son los más propensos a no concluir con los estudios, al igual que lo han señalado otros estudios (Rodríguez- Pineda, 2018; Rodríguez- Pineda y Zamora Araya, 2021).

Asimismo, con respecto a la carrera, edad y sexo del estudiantado, los datos muestran que entre mayor sea la edad a la que se ingrese a la universidad, el ser hombre y las personas que matriculan la carrera de administración de empresas, son las que tienen mayores probabilidades de abandonar el curso. Por otro lado, el profesorado con mayor experiencia y grado académico favorece la permanencia del estudiantado, por lo que se debe fomentar la capacitación constante del personal académico que imparte los cursos.

En lo que respecta a los algoritmos, aunque el poder predictivo fue discreto (entre un 62% y 64%), no es muy diferente al arrojado en otros estudios similares (Kilian et al., 2020; Maksimova et al., 2021; Muñoz-Camacho, 2018) y se debe tener en cuenta que sólo se cuenta con variables antes de iniciar el curso, las cuáles son las que suelen estar disponibles a la hora de plantear intervenciones tempranas. De acuerdo con el análisis realizado, no es posible afirmar que algún algoritmo sea mejor que otro ya que el ANOVA muestra que no existen diferencias significativas en cuanto al rendimiento predictivo, no obstante, la regresión logística fue el que presentó el mejor rendimiento en la métrica F1 (el de mayor porcentaje) y es el algoritmo más simple de interpretar de los que se evaluaron, por lo que se recomienda su uso para las personas tomadoras de decisión que, usualmente, no están familiarizadas con el uso de algoritmos de ML.

Finalmente, la predicción del abandono en el curso de Matemática General utilizando aprendizaje supervisado puede ser usada como herramienta para ayudar a seleccionar a las personas que serán beneficiarias de futuras intervenciones educativas, como aquellas que atiendan a grupos de interés para la universidad, como lo son aquellos que provienen de contextos socioeconómicos y educativos más desfavorables (como el estrato 3).

Por ejemplo, si el modelo de regresión logística se hubiera usado para identificar al estudiantado con mayores probabilidades de deserción en el curso de Matemática General que ingresaron en 2019, se hubiera determinado que el sexo y la edad tanto del profesorado como del estudiantado, el IDS, la beca, la carrera, el estrato, el financiamiento del colegio de procedencia, la nota en la PAA, la nota de colegio, la especialidad y grado académico docente son, entre todas las variables disponibles en los sistemas universitarios, las más relevantes para predecir el abandono. Con esta información, no solo se puede seleccionar al estudiantado que recibirá intervenciones, en función de su probabilidad de desertar, sino asignar a la o las personas docentes encargadas de ayudar en esas intervenciones con base en su perfil.

En la misma línea, supónganse que se desea plantear una intervención para trabajar una metodología que permita reforzar conocimientos previos en Matemática, pero por cuestiones metodológicas y de presupuesto solo se puede implementar con 15 estudiantes, pero hay 100 personas de estrato 3 interesadas en recibir la intervención, el algoritmo de aprendizaje supervisado de regresión logística podría usarse para seleccionar a las 15 personas, de entre las 100, que tienen mayor probabilidad de desertar y así éstas serían las que reciban la intervención.

Luego de la pandemia de COVID-19, todos los cursos del área de Matemática regresaron a la modalidad presencial, pero las secuelas de este periodo se evidencian en bajos rendimientos académicos y altos porcentaje de abandono, particularmente en cursos introductorios del área científico-matemática. Ante la falta de un sistema de alerta temprana para detectar el abandono que tenga acceso a datos en tiempo real, el modelo propuesto (en este caso el que se basa en la regresión logística) es una opción para que las personas tomadoras de decisiones puedan identificar, a tiempo, a los estudiantes con mayores probabilidades de desertar y así poder brindarles alternativas que permitan aumentar los índices de permanencia.

El objetivo de este estudio no consiste en señalar qué acciones aplicar o cuándo, sino ayudar a identificar a quienes aplicar en un contexto de restricciones al presupuesto universitario. Por

otra parte, los modelos propuestos, además de calcular la probabilidad de deserción por estudiante, brindan evidencia empírica sobre cuáles variables son las que mejor ayudan a predecir el abandono en el curso con uno de los rendimientos más bajos y altas índices de desincorporación y, así, prevenir en etapas tempranas la desincorporación del estudiantado. Finalmente, futuras investigaciones podrían incorporar nuevas variables, que sean susceptibles de recopilarse de forma masiva, para que puedan evaluar su capacidad para predecir el abandono y, eventualmente, mejorar el rendimiento de los algoritmos utilizados u otros que se deseen evaluar.

Referencias

Burkov, A. (2019). *The hundred-page Machine Learning book en español*. Andriy Burkov.

Calva, K., Flores, M., Porras, H., & Cabezas-Martínez, A. (2021). Modelo de predicción del rendimiento académico para el curso de nivelación de la Escuela Politécnica Nacional a partir de un modelo de aprendizaje supervisado. *Latin-American Journal of Computing*, *8*(2), 58–71. https://doi.org/10.5281/zenodo.5770905

Cao, C., Chicco, D., & Hoffman, M. M. (2020). *The MCC-F1 curve: A performance evaluation technique for binary classification*. https://doi.org/10.48550/arXiv.2006.11278

Castillo-Sánchez, M., Gamboa-Araya, R., & Hidalgo-Mora, R. (2020). Factores que influyen en la deserción y reprobación de estudiantes de un curso universitario de matemáticas. *Uniciencia, 34*(1), 219–245. http://hdl.handle.net/11056/20072

Contreras, L. E., Fuentes, H. J., y Rodríguez, J. I. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formación Universitaria,13* (5), 233–246. http://dx.doi.org/10.4067/S0718-50062020000500233

Hidalgo, R., Gamboa, R., y Castillo, M. (2019). Caracterización y posibles causas de la deserción estudiantil en el curso de Matemática General. Memorias del I Congreso Internacional de Ciencias Exactas y Naturales de la Universidad Nacional, Costa Rica, 1–9. https://www.researchgate.net/publication/334501920

Kilian, P., Loose, F., & Kelava, A. (2020). Predicting math student success in the initial phase of college with sparse information using approaches from statistical learning. *Frontiers in Education, 5*, 502698. https://doi.org/10.3389/feduc.2020.502698

Kuhn, M., & Johnson, K. (2016). *Applied predictive modeling*. Springer.Michigan

Kuz, A., & Morales, R. (2023). Ciencia de datos educativos y aprendizaje automático: Un caso de estudio sobre la deserción estudiantil universitaria en México. *Education in the Knowledge Society (EKS), 24,* e30080. https://doi.org/10.14201/eks.30080

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software.* https://doi.org/10.21105/joss.01903

López-Zambrano, J., Lara-Torralbo, J. A., Romero-Morales, C., *et al*. (2021). Early prediction of student learning performance through Data Mining: A systematic review. *Psicothem*a, 33(3), 456–465. https://doi.org/10.7334/psicothema2021.6

Maksimova, N., Pentel, A., y Dunajeva, O. (2021). *Predicting first-year computer science students drop-out with Machine Learning methods: A case study. Educating Engineers for Future Industrial Revolutions: Proceedings of the 23rd International Conference on Interactive Collaborative Learning (ICL2020),* Volume *2* 23, 719–726. https://doi.org/10.1007/978-3-030-68201-9_70

Mahesh, B. (2020). Machine Learning algorithms-a review. *International Journal of Science and Research (IJSR), 9*, 381–386. https://doi.org/10.21275/ART20203995

Muñoz-Camacho, S. V., Gallardo, T., Muñoz-Bravo, M., & Muñoz-Bravo, C. A. (2018). Probabilidad de deserción estudiantil en cursos de Matemáticas básicas en programas profesionales de la Universidad de los Andes-Venezuela. *Formación* Universitaria, 11(4), 33–42. http://dx.doi.org/10.4067/S0718-50062018000400033

R Core Team. (2022). R: A language and environment for statistical computing (Version 4.2.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rodríguez-Pineda, M. (2018). *De la reproducción social en la permanencia en la educación superior, caso de la Universidad Nacional de Costa Rica*. Congresos CLABES. https://revistas.utp.ac.pa/index.php/clabes/article/view/1948

Rodríguez-Pineda, M., & Zamora-Araya, J. A. (2021). Abandono temprano en estudiantes universitarios: Un estudio de cohorte sobre sus posibles causas. *Uniciencia, 35*(1), 19–37. http://dx.doi.org/10.15359/ru.35-1.2

Ruto, N. (2022). *How to get started with the boruta algorithm in Machine Learning*. https://towardsdatascience.com/boruta-explained-the-way-i-wish-someone-explained-it-to-me-4489d70e154a

Smith, J., y Gutiérrez, C. (2022). Una aplicación de aprendizaje automático (machine learning) en políticas públicas. Predicción de alerta temprana de deserción escolar en el sistema de educación pública de Chile. *Multidisciplinary Business Review,15*(1), 20–35. http://dx.doi.org/10.35692/07183992.15.1.4

Solís, M., Moreira, T., González, R., Fernández, T., & Hernández, M. (2018). Perspectives to predict dropout in university students with Machine Learning. 2018 *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 1–6. https://doi.org/10.1109/IWOBI.2018.8464191

Tinto, V. (1982). Defining dropout: A matter of perspective. *New Directions for Institutional Research, 1982* (36), 3–15. https://doi.org/10.1002/ir.37019823603

Universidad Nacional [UNA]. (2022). Procedimientos de admisión para ingreso a las carreras de grado de la Universidad Nacional. Universidad Nacional. http://documentos.una.ac.cr/handle/unadocs/1093

Zamora-Araya, J. A., & Villalobos-Madrigal, F. J. (2018). Factors associated with dropping out of the program for bachelor's and licentiate's degrees in mathematics teaching at the Universidad Nacional de Costa Rica (UNA): Evidence from the 2016 student cohort. Uniciencia, 32(2), 111–126. http://dx.doi.org/10.15359/ru.32-2.8