

Diseño de un modelo predictivo para generar alertas de deserción temprana en educación superior

Yudy Andrea Quintero¹, John Freddy Duitama²

¹ Estudiante de maestría en ingeniería, Facultad de Ingeniería, Universidad de Antioquia Medellín, Colombia
yudy.quintero@udea.edu.co.

² Docente departamento de ingeniería de sistemas, Facultad de Ingeniería, Universidad de Antioquia Medellín, Colombia.
john.duitama@udea.edu.co

PALABRAS CLAVE:

Deserción temprana; data mining; modelo predictivo; machine learning; learning analytics.

RESUMEN:

El fenómeno de la deserción es una problemática que aqueja a la mayoría de las instituciones de educación superior en el país, por esta razón ha sido ampliamente estudiado para determinar sus posibles causas e implementar acciones en pro de su disminución, para lo cual se ha hecho usos de diferentes técnicas y herramientas del análisis estadístico y computacional. La facultad de ingeniería no es ajena a dicho fenómeno pues observa que el nivel de deserción temprana promedio es de alrededor del 40%, por tal motivo se han realizado variadas investigaciones que han apuntado al estudio del fenómeno desde su caracterización e identificación de las causas que lo producen. Con este proyecto de investigación se buscó ir más allá de la caracterización del fenómeno y proponer una herramienta que identifique de manera temprana aquellos estudiantes que están en riesgo de desertar. Para lograr tal propósito se hizo uso de los datos históricos contenidos en las bases de datos de la universidad y la información disponible en la plataforma del Icfes y aplicando dos técnicas de *machine learning* como son redes neuronales artificiales RNA y *Xtreme gradient boosting* – XGBoost se entrenaron diferentes modelos. Se construyeron modelos para los tres primeros semestres, pues el propósito fue identificar la deserción temprana. El modelo de primer semestre contiene solo los datos de inscripción y para los modelos de segundo y tercer semestre se incluyeron variables de desempeño académico del estudiante. Encontrando que a medida que el estudiante avanza en su proceso de formación los modelos van logrando una mejor capacidad predictiva.

1. Introducción

Actualmente la deserción estudiantil en la educación superior es un problema que afecta a la mayoría de las universidades en toda América Latina, motivo por el cual a la fecha existe una gran cantidad de investigaciones al respecto de este fenómeno en donde se dan cuenta del gran número de estudiantes que no logran culminar satisfactoriamente con sus estudios universitarios y los costos sociales que esto les representa [1].

La Universidad de Antioquia no es ajena a este fenómeno y en particular la facultad de ingeniería en donde, según el estudio realizado por el grupo de investigación Ingeniería y Sociedad, en el marco del Observatorio de la vida académica en esta facultad, en las cohortes entre el 2005 y 2015 se encontró que la deserción acumulada en 10 semestres fue del 51%, otro hallazgo a destacar de este estudio es que la una gran proporción de los que desertaron lo hicieron en el primer semestre. Dicho estudio junto con otros realizados en la universidad y en la facultad [2], [3], [4], [5], han servido para dar claridad y explicar el problema de la deserción estudiantil en la Universidad de Antioquia y en la facultad de ingeniería, sin

embargo, estos han abordado el fenómeno a partir de su ocurrencia para la determinación de las causas que lo originan, es decir, se analiza el problema una vez que este ha ocurrido. No obstante, haciendo uso adecuado de la información disponible en las bases de datos institucionales y aplicando las técnicas de *machine learning* es posible construir un modelo con la capacidad de predecir la probabilidad que tiene un estudiante de desertar basado en un conjunto de características que este exhibe [6].

Learning Analytics (LA) o *Educational Data Mining* (EDM) es un área de investigación y desarrollo que, a partir de la recolección, monitoreo, análisis y modelamiento de la información obtenida de los estudiantes y su entorno de aprendizaje busca optimizar los procesos de enseñanza y aprendizaje [7]. Esta área de conocimiento, que surgió en los últimos años, ofrece diferentes tipos de soporte computacional para monitorear el comportamiento de los estudiantes a fin de hallar patrones y/o tendencias ocultas, además, de identificar correlaciones en los datos de carácter educativo para proveer información de valor a los profesores y administradores acerca de los estudiantes y el entorno en que estos se desenvuelven [6].

En el presente artículo se propone el diseño de un modelo predictivo basado en *Learning Analytics* que permite generar alertas de deserción temprana en la facultad de ingeniería en la Universidad de Antioquia, a partir de los datos disponibles en las bases de datos institucionales y los datos ofrecidos por la plataforma del Icfes.

2. Desarrollo del modelo

2.1. Definición de deserción

Existen varias definiciones para el concepto de deserción ya que el estudio de esta abarca una variedad de perspectivas, por lo cual existen diferentes tipos de deserción los cuales se ubican dentro de dos categorías [9]. La primera categoría establece la deserción en el tiempo, la cual a su vez puede ser precoz, que se presenta cuando el estudiante no alcanza a iniciar su proceso de formación, temprana que se presenta cuando el estudiante deserta en los primeros años de estudio y tardía que indica que el abandono se presenta cuando el estudiante llevaba un avance considerable en su proceso formativo. La otra categoría de deserción es la que se establece en el espacio; en esta clasificación se presentan: la deserción interna o del programa, la deserción institucional que es cuando el estudiante abandona la institución educativa y finalmente la deserción del sistema en la cual se considera como desertor a aquel que abandona definitivamente el proceso de formación [10] y [11]. Hay que tener en cuenta que el estudio de la deserción debe realizarse desde diferentes perspectivas de acuerdo con el investigador y la situación del ambiente en el que está el estudiante, por tal motivo las variables para incluir en el estudio dependen del punto de vista desde el cual se haga el análisis; esto es, individual, institucional y estatal o nacional [9]. Para fines de esta investigación se estableció que el tipo de deserción a analizar es la deserción en el tiempo en la facultad de ingeniería de la universidad de Antioquia, particularmente la deserción temprana, ajustándose a la definición establecida por la universidad: “desertores tempranos son aquellos estudiantes que permanecen en la universidad durante 3 semestres o menos”; esto es, se etiquetaron como desertores aquellos estudiantes que registraron matrícula durante solo 3 semestres o menos.

2.2. Descripción del conjunto de datos

El conjunto de datos utilizado proviene de los diferentes sistemas de información institucionales, **MARES** (Sistema de matrícula y registro académico) y **MOISES** (Sistema de Información de Inscripción y Selección de Estudiantes) en conjunto con la información que suministra El ICFES a través de la base de datos pública **DataIcfes**, la cual almacena la información asociada a los estudiantes que aplican a los diferentes exámenes, los resultados de desempeño de estos en las pruebas, su información socioeconómica, información asociada a la institución de educación media en la cual está matriculado, entre otros. [12]

El conjunto de datos resultante fue de 15187 registros, correspondientes a estudiantes que ingresaron a la universidad en el periodo comprendido entre años 2000-1 a 2017-2 y pertenecientes a los diferentes programas de ingeniería ofertados por la facultad en la sede principal de la universidad. Del total de

registros 5,950 fueron etiquetados como desertores tempranos, es decir, que la tasa de deserción temprana promedio durante todo el periodo evaluado fue de alrededor del 39%.

2.3. Selección de variables

Muchas discusiones se han dado alrededor de la deserción estudiantil y a pesar de que no se tiene una definición unificada sobre este fenómeno [13], existe un consenso en cuanto a que este debe ser explicado por diferentes grupos de variables las cuales deben ser de carácter: socio económicas, individuales, institucionales y académicas [11]. De acuerdo con este planteamiento del conjunto de datos obtenido, se seleccionaron las variables que hacen parte de cada una de las categorías de variables que influyen en la deserción. Dichas variables con su respectiva categoría se observan en la tabla 1.

Tabla 1. Variables seleccionadas

Categoría	Nombre de la variable
Socio económica	Educación de la madre, educación del padre, ocupación de la madre, ocupación del padre, estrato, situación laboral, personas a cargo, número de hermanos, Región, si proviene de una ciudad capital, si proviene de una cuida intermedia
Individual	Edad, genero, estado civil, discapacidades, tiempo para ingreso a la universidad después de terminado el colegio.
Institucional	Programa
Académica	Puntaje pruebas de admisión, experiencia previa en la educación superior, ingreso por beca. Créditos matriculados, aprobados, cancelado o reprobados. Promedio semestre, colegio del cual egresó, categoría colegio

Estas son las variables de entrada del modelo o variables dependientes y la variable objetivo es la condición de “desertor”, 1 si el estudiante desertó de manera temprana o 0 si el estudiante no desertó.

2.4. Experimento

Dado que el tipo de deserción que se estudió fue la deserción temprana y de acuerdo con la definición establecida por la UDEA que indica que esta es la que se da en los 3 primeros semestres [13], se optó por entrenar 3 modelos diferentes: un modelo 0 que se da en el momento en que el estudiante ingresa a la universidad y matricula su primer semestre, en este se contó con los datos de inscripción en el proceso de matrícula y la información recuperada del ICFES. El modelo 1 se da en el momento en que el estudiante logró permanecer un semestre en la universidad y se matricula para el segundo semestre, en este caso se cuenta con la misma información del modelo 0 y adicional se incluyeron los resultados del desempeño académico obtenido durante el primer semestre. Finalmente, para el modelo 2, se cuenta con toda la información del modelo 1 y adicionalmente la información de desempeño académico durante el segundo semestre. Para cada modelo se utilizaron dos técnicas diferentes las cuales tienen una gran capacidad predictiva como lo son: las Redes neuronales artificiales RNA y el eXtreme Gradient Boosting XGBoots. Se entrenaron los modelos usando ambas técnicas, modificando los hiperparámetros de cada uno, de manera que se lograra obtener el mejor de rendimiento de cada técnica usada. Los resultados de los entrenamientos se observan en la tabla 2, la cual muestra la precisión obtenida por cada una de las técnicas para predecir la deserción temprana en los diferentes momentos evaluados. Se puede ver que a medida que se avanza en el programa de formación se puede predecir con mayor precisión si el estudiante se encuentra en riesgo de desertar. De los resultados expuestos también se logra evidenciar que el XGBoots logró un mejor desempeño frente a las RNA para el conjunto de datos trabajado.

Tabla 2. Resultado de los modelos entrenados

	Técnica	Media de <i>accuracy</i> en entrenamiento	Media de <i>accuracy</i> en prueba
Modelo_0	XGBoots	75,35 %	73,11 %
	RNA	72,80 %	71,93 %
Modelo_1	XGBoots	89,23 %	87,81 %
	RNA	84,17 %	82,14 %
Modelo_2	XGBoots	91,72%	91,23%
	RNA	91,48 %	90,20 %

3. Conclusiones

En este trabajo se propone un modelo con la capacidad para predecir el riesgo de deserción temprana en la facultad de ingeniería usando una base de datos que incluye variables tanto socio económicas como académicas e individuales, sin embargo, se contó con pocas variables que hicieran parte del ámbito institucional pues de esta categoría solo se incluyó el programa y según lo expone la teoría dado que la deserción es un fenómeno multicausal se deben contar con la mayor cantidad de variables posibles de todas las categorías a fin lograr una mejor explicación del fenómeno. A pesar de que el resultado se considera aceptable, pues se logró una precisión superior al 70% en el modelo 0. Si se hubiera contado con una mayor cantidad de variables se hubiera podido lograr un mejor desempeño de los modelos.

A medida que se avanza en los semestres la precisión en la predicción de la deserción fue mejorando, esto a medida que se fueron incluyendo las variables que dan cuenta del desempeño académico del estudiante por lo cual se observa que una vez que este ha logrado superar el primer semestre, su desempeño académico se vuelve determinante a la hora de tomar la decisión de abandonar sus estudios en el segundo y tercer semestre.

Contar con una herramienta que genere alertas tempranas de deserción es de gran beneficio para la facultad de ingeniería puesto que arroja luces para que desde el área administrativa se tome acciones en pro de la permanencia de los estudiantes.

REFERENCIAS

Data Availability Statement

Referencias

- [1] S. Cortés, P. Álvarez, M. L. Ortiz and L. C. Collado, “CI Deserción universitaria: La epidemia que aqueja a los sistemas de educación superior”. REVISTA PERSPECTIVA, Vol. 20(1), pp 13-25, 2019
- [2] E. Castaño, and J. Vásquez, “Deserción estudiantil universitaria una aplicación de modelos de duración”. Lecturas de Economía, 60(60), pp. 39–65, 2004.
- [3] C. M. Parra, L. F. Mejía, A. Valencia, E. Castañeda, G. Restrepo, O. Úsuga, and R. Mendoza, “Rendimiento académico de los estudiantes de primer semestre de pregrado de la Facultad de Ingeniería de la Universidad de Antioquia: cohorte 2012-2”, Ingeniería y sociedad, Vol. 6, pp. 2–10, 2013.
- [4] E. Castaño, “Rendimiento académico de los estudiantes en el primer semestre: Facultad de Ingeniería. Cohortes 2016-1 y 2015-1”, Revista Ingeniería y Sociedad, Vol. 11, pp. 27–33, 2016.

- [5] M. A. Ciro Úsuga, “La deserción estudiantil universitaria: análisis relacional del fenómeno en la Universidad de Antioquia para la cohorte 2009- 1”, Departamento de Sociología, Universidad de Antioquia, Medellín, 2017.
- [6] S. Bingulac, “On the compatibility of adaptive controllers,” in 4th Annu. Allerton Conf. Circuit and Systems Theory, New York, USA, 1994, pp. 8–16.
- [7] V. D. Gil Vera, “Learning Analytics and Scholar Dropout. A Predictive Model”, Middle-East Journal of Scientific Research vol. 25, pp. 1414–1419, January 2017.
- [8] [D.K. Mah, “Learning Analytics and Digital Badges: Potential Impact on Student Retention in Higher Education”, Technology, Knowledge and Learning, Vol. 21(3), p 285–305, 2016.
- [9] I. M. Quintero Análisis de las causas de deserción universitaria. [online]. Disponible en: <https://repository.unad.edu.co/handle/10596/6253>.
- [10] N. Murcia and P. J. Ramírez, (2015). “Motivos de deserción estudiantil en programas virtuales de posgrado: revisión de caso y consideraciones desde el mercadeo educativo y el mercadeo relacional para los programas de retención”, Revista de Educación a Distancia (RED), no 45, 2016
- [11] C. Ruiz, D. Muriel, J. Gallego, J., E. Castaño, S. Gallón Gómez, K. Portilla and J. Vásquez, “Deserción estudiantil en la educación superior colombiana”, Ministerio de Educación Nacional, 2009.
- [12] J. A. Estévez, J. Castro, H. GRANOBLES, “La educación virtual en Colombia: exposición de modelos de deserción”. Apertura, vol. 7, no 1, pp. 1-10, 2015.
- [13] Instituto Colombiano para la Evaluación de la Educación, ICFES, Acceso a bases de datos y diccionarios, 2020. [Online]. Available: <https://www.icfes.gov.co/investigadores-y-estudiantes-posgrado/acceso-a-bases-de-datos>. Accessed on: Aug. 26, 2021.
- [14] Universidad de Antioquia, “Plan de acción de Acción institucional 2018- 2021,”, 2019. [Online]. Available:<https://www.udea.edu.co/wps/portal/udea/web/inicio/institucional/direccionamiento-estrategico/plan-accion-institucional>. Accessed on: Aug. 27, 2021.