Encuentro Internacional de Educación en Ingeniería ACOFI
GESTIÓN, CALIDAD Y DESARROLLO EN LAS FACULTADES DE INGENIERÍA
Cartagena de Indias, Colombia
18 al 21 de septiembre de 2018

MODELO DE PREDICCIÓN PARA LA DESERCIÓN TEMPRANA EN LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD DE LA SALLE

Heriberto Felizzola Jiménez, Yamile Adriana Jaime Arias, Ana María Castillo Pastrana, Fidelina Villa Pedroza

Universidad de La Salle
Bogotá, Colombia

Resumen

La deserción estudiantil es un fenómeno complejo que involucra diversos factores en los ámbitos sociales, económicos, familiares, psicologícos y académicos del estudiante. Estudios previos del *Ministerio de Educación Nacional*, MEN, indican que algunos de los factores determinantes en la deserción son: estrato, sexo, nivel educativo de los padres, ingresos económicos de la familia, clasificación según el SISBÉN, número de personas que componen el núcleo familiar, resultados de las Pruebas de Estado Saber 11° y ocupación del joven.

Dada la complejidad del problema y el gran impacto que esto genera a nivel social, las universidades diseñan estrategias de intervención que permitan disminuir la tasa de deserción. El inconveniente es que muchas de estas estrategias carecen de efectividad, ya que, no tienen en cuenta que las causas varían en cada caso. Por otro lado, se necesita información confiable que permita caracterizar la población para identificar posibles casos de deserción antes que ocurran, con lo cual se puedan tomar acciones preventivas que permitan disminuir la tasa de deserción.

En este sentido, el propósito de la investigación es diseñar un modelo de clasificación para la deserción temprana en la Facultad de Ingeniería de la Universidad de la Salle, a través de la aplicación de la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*)

Se ha realizado una revisión de la literatura de 1982 a 2017, en la cual se analizan las aplicaciones de *machine learning* y *data mining* para abordar la problemática con métodos como *decision trees, artificial neural networks, support vector machines, naive bayes, uniform random, k nearest neighbor, logistic regression*, entre otros;

Palabras clave: deserción estudiantil; minería de datos; aprendizaje automático

Abstract

Student desertion is a complex phenomenon that involves several factors in the social, economic, family, psychological and academic fields of the student. Previous studies of the Ministry of National Education, MEN, indicate that some of the determining factors in the desertion are: economic status, gender, educational level of the parents, income of the family, classification according to the SISBÉN, number of people that make up the family of the nucleus, the results of the Saber State Tests 11th and the student's occupation.

Given the complexity of the problem and the great impact that this generates at a social level, the universities design intervention strategies that allow reducing the school dropout rate. The drawback is that many of these strategies lack effectiveness, since they do not take into account that the causes vary in each case. On the other hand, reliable information is needed to characterize the population in order to identify possible cases of desertion before they occur, with which preventive actions can be taken to reduce the student dropout rate.

In this sense, the objective of this research is to design a classification model for premature abandonment in the Faculty of Engineering of the Universidad de la Salle, through the application of the CRISP-DM methodology (Cross Standard Process for Data Mining)

A review of the literature from 1982 to 2017 has been carried out, in which the applications of machine learning and data mining are analyzed to approach the problem with methods such as decision trees, artificial neural networks, support vector machines, naive bays, uniform random, k nearest neighbor, logistic regression, among others;

Keywords*: student desertion; data mining: machine learning*

1. Introducción

Uno de los problemas esenciales que presenta el sistema de educación superior en Colombia, son las altas tasas de deserción estudiantil, especialmente en pregrado. El número de alumnos que logra terminar sus estudios es bajo, debido a que, la mayoría de éstos los abandonan, principalmente en los primeros semestres.

El informe sobre educación superior en América Latina y el Caribe, presentado por la UNESCO (2016), estudia la deserción en esta área geográfica, promediando las áreas de conocimiento con las tasas de deserción de los países de estas regiones, teniendo como resultado que: el área Salud tiene una tasa de graduación de 54.2%, la administración y comercio del 49.6%, el derecho del 49%, la educación del 48.8%, las ciencias sociales del 47.4%, la agricultura del 41.9%, el arte y la arquitectura del 40.8%, la tecnología y la ingeniería 38, 5 %, la ciencia básica de 36.8% y el área de humanidades de 23.1%.

La deserción tiene un lugar esencial en la sociedad ya que las pérdidas que representan son altas en los ámbitos financiero y social, para el individuo, las familias, las instituciones de educación superior y la sociedad. Estudiar el problema de la deserción también proporciona políticas de control efectivas para proporcionar un aumento en la cobertura en educación con calidad y equidad. Usando técnicas de aprendizaje automático y minería de datos se logra obtener predicciones a través del estudio de patrones ocultos en sistemas de información (base de datos), permiten visualizar resultados para que puedan ser leídos y entendidos de manera inmediata, simple y efectiva; útil para tomar decisiones estratégicas para la solución del problema.

De esta manera, el programa de Ingeniería Industrial busca analizar la problemática de la deserción estudiantil en la Facultad de Ingeniería de la Universidad de La Salle, a través de la creación de un modelo con herramientas de minería de datos, el cual permita: 1) analizar los factores que afectan la deserción dentro de la facultad, y 2) predecir la probabilidad de deserción que pueda tener un estudiante. Esto, podría generar conocimientos que la universidad pueda utilizar, para crear estrategias que promuevan la retención estudiantil, y, por lo tanto, lograr mayores tasas de graduación y menores de deserción.

2. Marco Teórico

El marco teórico de esta investigación se segmenta en dos componentes: (1) la contextualización del concepto de deserción y variables asociadas a ella, y, (2) las técnicas existentes de *machine learning* y *data mining* en los estudios sobre la deserción.

El Ministerio de Educación Nacional de Colombia (2009), define la deserción como *"una situación a la que se enfrenta un estudiante cuando aspira y no logra concluir su proyecto educativo, considerándose como desertor a aquel individuo que siendo estudiante de una institución de educación superior no presenta actividad académica durante dos semestres académicos consecutivos, lo cual equivale a un año de inactividad académica".*

Según Tinto (1982), la deserción tiene en cuenta un conjunto de variables que hacen determinantes el hecho de que el estudiante se gradúe o abandone sus estudios. Las principales variables son: las **individuales**, como edad, género, estado civil, entorno familiar, calamidad y problemas de salud, integración social, incompatibilidad horaria con actividades extra académicas, expectativas no satisfechas, embarazos no deseados; las **académicas,** como orientación profesional, tipo de colegio, rendimiento académico, calidad del programa, métodos de estudio, resultado en el examen de ingreso, insatisfacción con el programa u otros factores, número de materias; las **institucionales**, como normalidad académica, becas y formas de financiamiento, recursos universitarios, orden público, entorno político, nivel de interacción personal con los profesores y estudiantes, apoyo académico, apoyo psicológico; y las **socioeconómicas**, como estrato, situación laboral del estudiante y de los padres e ingresos, dependencia económica, personas a cargo, nivel educativo de los padres, entorno macroeconómico del país, localización de la vivienda.

Se pueden presentar dos tipos de deserción, una según el tiempo y la otra teniendo en cuenta el espacio. La deserción con respecto al tiempo se clasifica como abandono prematuro, deserción temprana, deserción tardía; y la deserción espacial se divide en deserción institucional y el programa interno o académico de deserción. La deserción prematura es aquella en la que el alumno ha sido admitido en la Universidad, pero no está inscrito. La deserción temprana es el abandono de los estudios por parte del individuo en los primeros semestres de la carrera. La deserción tardía es el abandono de los estudios en los últimos semestres de la carrera (Giovagnoli, 2001; Tinto, 1982).

La deserción institucional ocurre cuando el estudiante se retira de la Universidad para inscribirse en otra o desertar permanentemente. El programa interno o académico de deserción consiste en que el alumno haya decidido cambiar de carrera o programa a otro que ofrezca la misma Universidad.
Se ha buscado solucionar la problemática de la deserción universitaria a través de herramientas que permiten predecir las posibles características de un estudiante desertor junto con la probabilidad de que abandone sus estudios. Las técnicas de predicción más utilizadas en este contexto son las de *Machine Learning* y Minería de Datos.

2.1. Técnicas de aprendizaje automático y minería de datos

El aprendizaje automático es la programación de computadoras con algunos parámetros para optimizar un criterio de rendimiento utilizando datos de entrenamiento o conocimientos previos. El modelo puede ser predictivo para hacer predicciones en el futuro, o descriptivo para obtener conocimiento de los datos o ambos (Ethem, 2014).

Hay una gran cantidad de algoritmos de aprendizaje automáticos. Autores como Ethem (2014), Kotsiantis (2007), Zhang & Tsai (2007) clasifican las técnicas según el enfoque del proceso de aprendizaje, las cuales son: supervisada, no supervisada, semi-supervisada y refuerzo del aprendizaje.

Kotsiantis (2007) define el aprendizaje supervisado como un modelo conciso de la distribución de las etiquetas de clase con respecto a las características del predictor. El clasificador que se utiliza para asignar las etiquetas de clase a las instancias de prueba donde se conocen las características de los valores del predictor, pero se desconoce el valor de la etiqueta de la clase. Por otro lado, el aprendizaje no supervisado consiste en determinar modelos que se centran principalmente en la búsqueda de patrones ocultos en los datos, sin tener un conjunto de entrenamiento.

El aprendizaje semi-supervisado, según Chapelle, Scholkopf, & Zien, Eds. (2009), reside en crear un modelo a partir de un sistema de capacitación con información faltante, donde se aprende y se elimina el resultado con datos incompletos. Finalmente, Sutton & Barto (1998) describien el aprendizaje reforzado como un modelo de análisis matemático o estadístico para aprender basado en la retroalimentación externa dada por un cuerpo de pensamiento o el medio ambiente.

La minería de datos es "el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de forma automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un contexto determinado".

Las técnicas de minería de datos pueden usarse en diferentes contextos de acuerdo con el método que desee aplicar. Estos métodos son clasificación, análisis de asociación y agrupamiento (Tan, Steinbach, & Kumar, 2005). La clasificación, se encuentra estrechamente relacionada con el aprendizaje supervisado. Por otro lado, entre las técnicas relacionadas con aprendizaje no supervisado se encuentran: Análisis de asociación y agrupamiento. El primero, es útil para descubrir relaciones interesantes escondidas en grandes conjuntos de datos. Las relaciones descubiertas se pueden representar en forma de reglas de asociación, que son una expresión de la implicación de la forma $x \rightarrow y$. Mientras que el agrupamiento divide el conjunto de datos en grupos que son significativos. Si los grupos significativos son los objetivos, entonces las agrupaciones deben capturar la estructura natural de los datos.

Dentro de los métodos de clasificación y técnicas de aprendizaje supervisado se encuentran: redes neuronales artificiales, árboles de decisión, algoritmos bayesianos, regresión logística, máquinas de soporte vectorial, búsqueda de vecinos cercanos, o técnicas de ensamble de estos algoritmos.

3. Metodología

Las actividades para desarrollar esta investigación incluyen: (1) Realización de una revisión de la literatura en Colombia y a nivel mundial. (2) Análisis de la literatura para identificar las variables de estudio. (3) Recolección de los datos de las variables de estudio para los estudiantes de Ingeniería, mediante el uso de los Sistemas de Información, académico, financiero, de bienestar universitario, y de registro y admisiones, de la Universidad de la Salle. (4) Creación de una base de datos del modelo relacional. (5) Análisis exploratorio, aplicando estadística descriptiva, de las variables de estudio recolectadas. (6) Construcción del modelo de predicción utilizando técnicas de minería de datos. (7) Evaluación del modelo.

4. Resultados

Se han obtenido resultados en cuanto a la revisión de la literatura de la problemática de la deserción universitaria a nivel mundial y en Colombia. Con una revisión de cincuenta artículos, de los cuales se seleccionaron diez por ser los más más relevantes, en cuanto a las variables utilizadas, la evaluación de diferentes técnicas y los resultados obtenidos en cuanto a la capacidad predictiva de los modelos desarrollados para identificar casos de deserción. Entre los artículos destacados se encuentran:

Orea, Vargas, & Alonso (2010) buscaron predecir la deserción escolar en la Universidad Tecnológica de Izúcar de Matamoros. Como resultados encontraron que las causas primordiales de deserción según el modelo son la edad, los ingresos familiares y el nivel de inglés. Utilizaron como herramientas de minería de datos el algoritmo de árboles de clasificación C4.5 y el algoritmo de los k vecinos más cercanos, obteniendo la mejor precisión de 98,98% con el algoritmo C4.5.

Lin, Imbrie, & Reid (2009) compararon cinco modelos, estadísticos y de minería de datos, para analizar la retención a partir de factores cognitivos y / o no cognitivos. Utilizaron como herramientas de minería de datos las redes neuronales con una precisión del 71,7%.

Delen (2010) se enfocó en los factores de retención basado en los estudios de marketing y minería de datos "*churn analysis*", tomando cinco años de datos históricos. Utilizaron como herramientas de minería de datos las máquinas de soporte vectorial, redes neuronales artificiales, árboles de decisión (C5) y regresión logística; dando como resultado un promedio de capacidad predictiva del 80%, identificando que los factores académicos y financieros tienen mayor relevancia en el modelo.

Y. Zhang, Zhang, Oussena, Clark, & Kim (2010) construyeron un sistema de minería de datos que cubre el punto de vista académico, y el análisis de deserción con las variables que propone Tinto (1982). Recolectaron 3 años de registro de la universidad con 5,458 estudiantes. Utilizaron como herramientas de minería de datos el algoritmo de Naïve Bayes, máquina de soporte vectorial, árbol de decisión; siendo el de Naiive Bayes el de mejor precisión (89,5%).
Yu, Digangi, Jannasch-Pennell, & Kaprolet, (2010) ilustran cómo las técnicas de minería de datos pueden ser aplicado para estudiar los factores que afectan la retención de estudiantes universitarios. Utilizaron como herramientas de minería de datos técnicas de clasificación, regresión adaptativa multivariada (MARS), redes neuronales, obteniendo la mejor precisión con MARS del 67,4%.

Formia, Lanzarini, & Hasperue (2013) desarrollan un modelo que predice la probabilidad de deserción de los estudiantes de la Universidad Nacional de Río Negro (UNRN), y en particular en la Sede Atlántica desde la Licenciatura en Sistemas. Utilizaron como herramientas de minería de datos el algoritmo C4.5 con una precisión del 71,65%.

Jia & Mareboyana (2013) crearon algoritmos que se aplican para monitorear la retención de estudiantes de pregrado utilizando datos de estudiantes. El estudio también hizo algunas mejoras a los algoritmos de clasificación como árbol de decisión, máquinas de soporte vectorial (SVM) y redes neuronales, obteniendo la mejor precisión de 94,16% con las redes neuronales.

Thammasiri, Delen, Meesad, & Kasap (2014) compararon diferentes técnicas de balanceo de datos para mejorar la precisión de la predicción de la clase minoritaria, manteniendo un rendimiento de clasificación general satisfactoria. Utilizaron como herramientas de minería de datos las redes neuronales artificiales, máquinas de soporte vectorial, árboles de decisiones y regresión logística, teniendo una precisión del 90,24%.

Raju & Schumacker (2015) utiliza los datos de los primeros años de estudio disponibles para construir las técnicas de minería de datos, y así encontrar las características importantes asociados con la graduación con 22,099 observaciones totales en el conjunto de datos. Utilizaron como herramientas de minería de datos: redes neuronales, árboles de decisión y regresión logística, generando una precisión en la predicción del 70%.

Heredia, Amaya, & Barrientos (2015) muestra la construcción de un modelo predictivo de deserción escolar, caracterizando a estudiantes de la Universidad Simón Bolívar para predecir la probabilidad que un estudiante abandone su programa académico. Utilizaron como herramientas de minería de datos el algoritmo ID3, el algoritmo C4.5 y árboles de decisión, obteniendo una precisión del 92,9%.

5. Conclusiones

En la revisión de la literatura se observó que la técnica más utilizada es los árboles de decisión, dado que, esta genera las características más significativas que afectan el nivel de deserción de un estudiante. Dentro de las técnicas que generan un mayor porcentaje de precisión se encuentran las redes neuronales y los árboles de decisión. De otra parte, realizar ensambles no necesariamente proporciona un mayor índice de precisión en los modelos.

En la práctica, tratar de crear soluciones a la problemática de la deserción estudiantil se ha vuelto un tema que ha tomado fuerza en los últimos años, dado que, los países quieren tener un mejor nivel de educación, pero necesitan saber cuáles son las variables que tienen mayor incidencia, ya que, esto permite atacar el problema de forma directa y efectiva. En este sentido, las variables que más incidieron en la deserción estudiantil fueron las académicas y económicas.

6. Bibliografía

- Chapelle, O., Scholkopf, B., & Zien, Eds., A. (2009). Semi-Supervised Learning. *IEEE Transactions on Neural Networks*, *20*(3), 542–542.
- Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506.
- Ethem, A. (2014). *Introduction to machine learning. MIT press.* MIT Press. https://doi.org/10.1007/978-1-62703-748-8_7
- Formia, S., Lanzarini, L. C., & Hasperue, W. (2013). Caracterización de la deserción universitaria en la UNRN utilizando minería de datos. un caso de estudio. *Revista Iberoamericana de Tecnología En Educación y Educación En Tecnología*, *11*, 92–98.
- Giovagnoli, P. I. (2001). Determinantes de la deserción y graduación universitaria.
- Heredia, D., Amaya, Y., & Barrientos, E. (2015). Student Dropout Predictive Model Using Data Mining Techniques. *IEEE Latin America Transactions*, *13*(9), 3127–3134.
- Jia, J.-W., & Mareboyana, M. (2013). Machine Learning Algorithms and Predictive Models for Undergraduate Student Retention. *Proceedings of the World Congress on Engineering and Computer Science*, *11*, 23–25.
- Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. In *Emerging artificial intelligence applications in computer engineering* (pp. 3–24).
- Lin, J. J. J., Imbrie, P. K., & Reid, K. J. (2009). Student Retention Modelling: An Evaluation of Different Methods and their Impact on Prediction Results. *Proceedings of the Research in Engineering Education Symposium*, 1–6.
- Ministerio de Educación Nacional de Colombia. (2009). *Deserción estudiantil en la*

educación superior colombiana. Metodología de seguimiento, diagnóstico y elementos para su prevención.
- Orea, S. V., Vargas, A. S., & Alonso, M. G. (2010). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. *Recursos Digitales Para La Educacion y La Cultura*, 33–39.
- Raju, D., & Schumacker, R. (2015). Exploring Student Characteristics of Retention That Lead To Graduation in. *Journal of College Student Retention: Research, Theory & Practice*, *16*(4), 563–591.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning : an introduction*. MIT Press.
- Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Pearson Addison Wesley.
- Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, *41*(2), 321–330.
- Tinto, V. (1982). Limits of Theory and Practice in Student Attrition. *The Journal of Higher Education*, *53*(6), 687–700.
- UNESCO. (2016). *Informe sobre la Educación Superior en América Latina y El Caribe.* Caracas, Venezuela.
- Yu, C. H., Digangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year. *Journal of Data Science*, *8*, 307–325.
- Zhang, D., & Tsai, J. J.-P. (2007). *Advances in machine learning applications in software engineering*. Idea Group Pub.
- Zhang, Y., Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010). Using data mining to improve student retention in higher education: a case study. *IN INTERNATIONAL CONERENCE ON ENTERPRISE INFORMATION SYSTEMS*.

7. Sobre los autores

- **Heriberto Felizzola Jiménez**: Ingeniero Industrial, Master en Ingeniería Industrial. Profesor Asistente, healfelizzola@unisalle.edu.co
- **Yamile Adriana Jaime Arias**: Ingeniero Industrial, Master en Ingeniería Industrial. Profesor Asistente, yajaime@unisalle.edu.co
- **Ana María Castillo Pastrana**: Estudiante de Ingeniería Industrial. acastillo89@unisalle.edu.co.
- **Fidelina Villa Pedroza**: Estudiante de Ingeniería Industrial. fvilla58@unisalle.edu.co