V CLABES

QUINTA CONFERENCIA LATINOAMERICANA SOBRE EL ABANDONO EN LA EDUCACIÓN SUPERIOR

IMPLEMENTACIÓN DE MODELOS DE MINERÍA DE DATOS PARA LA DEFINICIÓN DE TENDENCIAS DE DESERCIÓN Y PERMANENCIA EN LA UNIVERSIDAD NACIONAL DE COLOMBIA

Línea Temática: Factores asociados al abandono, Tipos y perfiles de abandono, Factores asociados al abandono, Modelos y/o métodos de medición de las causas asociadas al abandono.

LÓPEZ GUARÍN, Camilo Ernesto
GALLEGO VEGA, Luis Eduardo
CASADIEGO, María Angélica
Universidad Nacional de Colombia - COLOMBIA
e-mail: sis_acompa_nal@unal.edu.co

Resumen. Colombia es un país de conflicto en el cual se generan problemáticas ambientales, sociales y familiares que dificultan que los estudiantes universitarios finalicen su proyecto académico en el tiempo previsto. Considerando, además, que las transformaciones, los cambios sociales y las modificaciones en la normativa institucional inciden de manera significativa en los índices de permanencia se evidenció la necesidad de identificar e intervenir los factores que afectan dichos índices y afirmar un modelo de acompañamiento académico que permita mantenerlos en rangos aceptables dentro de las instituciones de educación superior del país.
Por tal motivo, la Universidad Nacional de Colombia se planteó la creación de un modelo basado en técnicas de minería de datos para fortalecer la identificación de tendencias en torno a factores que influyen en el desempeño académico de los estudiantes. La ponencia presentará como, mediante el uso de estas técnicas conocidas por su valor predictivo e interpretabilidad (J48/C4.5, un árbol de decisión; Naïve Bayes, un clasificador Bayesiano, y regresión logística), c., se puede construir un modelo predictivo que permita identificar a los estudiantes que perderían la calidad de estudiante en su primera matrícula por bajo desempeño académico, facilitando a las instancias, tanto académicas como de bienestar, implementar acciones que les permitan actuar de manera oportuna sobre los factores que pueden afectar la permanencia de los estudiantes. Así mismo, el documento resalta cómo la implementación de estos modelos puede facilitar la creación de perfiles de estudiantes con riesgo académico, lo cual permitirá a las instituciones generar estrategias que actúen sobre las necesidades reales de los estudiantes universitarios para así disminuir la deserción, facilitar la permanencia y egreso.

Descriptores o Palabras Clave: Deserción, Permanencia, Factores Asociados, Educación Superior, Minería de Datos.

1 Introducción

De manera permanente las Universidades emprenden acciones de acompañamiento con el fin de disminuir el abandono estudiantil. Para que dichas acciones cumplan su objetivo se requiere generar mecanismos cada vez más eficaces que permitan detectar oportunamente a los estudiantes que podrían desertar, esto con el fin de emprender acciones que generen condiciones de permanencia y egreso cada vez más favorables.

En el contexto colombiano se ha detectado que "el aumento de la cobertura en el país elevó también la vulnerabilidad académica de los estudiantes nuevos, mostrando lo que podría denominarse un proceso de segunda selección durante la trayectoria" (Pinto, Durán, Pérez, Reverón, & Rodríguez, 2007:11). De acuerdo al Sistema para la Prevención de la Deserción de la educación Superior (SPADIES, 2015) del Ministerio de la Educación Nacional el 15% de los estudiantes en programas universitarios desertan habiendo realizado el primer semestre y después de 10 semestres la deserción acumulada para los estudiantes universitarios es cerca del 45%.

Ante estas cifras y considerando que en Colombia solo cerca del 46 % de la población entre los 17 y los 21 años accede a la Educación Superior, la Universidad Nacional de Colombia, como la principal institución pública de Educación Superior del país, vio imperativo identificar los factores que inciden de manera significativa en los índices de permanencia y deserción por lo que se hizo necesario identificar e intervenir dichos índices y afirmar un modelo de acompañamiento académico que permita mantenerlos en rangos aceptables dentro de la institución.

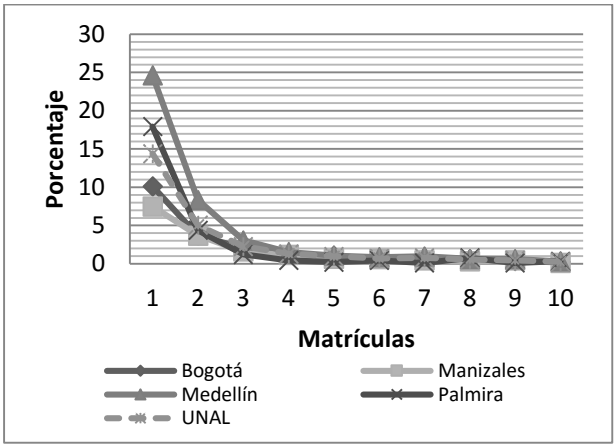
Para iniciar esa labor la Universidad tuvo en cuenta la naturaleza multifactorial del fenómeno de la deserción, que en el contexto colombiano está relacionada con las problemáticas ambientales, políticas, sociales y familiares que dificultan que los estudiantes universitarios finalicen sus estudios. Condiciones a las que se suman, las modificaciones en la normativa institucional.

La Universidad implementó una reforma académica en el año 2009 que implicó cambios en la gestión curricular, la cual se centró en la flexibilidad y el fortalecimiento de la autonomía de los estudiantes. La posibilidad que ellos escogieran sus trayectos de formación trajo consigo una reflexión profunda en la institución acerca de las dificultades de adaptación de los estudiantes que están ingresando cada día más jóvenes y nuevos desafíos que se le plantean debido a las condiciones culturales y académicas con las que ingresan a la Educación Superior.

Con el fin de generar mecanismos efectivos de detección de estudiantes en riesgo de deserción, la Universidad Nacional de Colombia emprendió el ejercicio de construir, mediante el uso de técnicas de minería de datos, un modelo predictivo que permitiera identificar a los estudiantes que perderían su vinculación por bajo desempeño académico.

Para la construcción del modelo se tomó como base el análisis que realizó la Vicerrectoría Académica de la Universidad Nacional de Colombia en el que se identificó que la institución tiene una deserción académica del 28,4%, la cual se concentra especialmente en el primer semestre con un porcentaje del 14,4 % (Fig. 1):

Figura 1. Deserción Académica - UNAL



De acuerdo a lo anterior, en la Universidad existe una alta deserción temprana la cual, si se logra mitigar podría disminuir la deserción acumulada. Al respecto, cabe destacar que la desvinculación, sea temprana o forzosa, aumenta la probabilidad de abandono de los estudios. Es decir, que se si se detectan los factores que provocan la desvinculación se podría entender la naturaleza de la deserción en la Universidad. Con esta hipótesis se planteó el desarrollo de un modelo de predicción basado en técnicas de minería de datos que a partir de la información del estudiante clasifica la variable de interés: si fue desvinculado o no de la Universidad durante su primera matrícula. Estas fuentes corresponden a factores académicos, socioeconómicos y sociodemográficos que pueden ser detectados dado que los estudiantes suministran esta información desde el proceso admisión y matrícula a la Universidad.

Ahora bien, el objetivo del estudio es identificar las tendencias de los estudiantes en riesgo de abandono en el primer semestre a partir de la información de entrada a la Universidad. Así como determinar el tipo de modelo que sería más efectivo para realizar esta tarea predictiva en la Universidad, lo que presupone decantar las fuentes de información con el fin de seleccionar los datos que serían más adecuados para definir las tendencias de deserción estudiantil.

Se espera con esta propuesta encontrar las herramientas con las cuales se podría obtener los más altos índices de predicción de la deserción para construir una base que permita elaborar un perfil del desertor en la Universidad Nacional de Colombia.

2 Metodología

El uso de técnicas de minería de datos ha aumentado en los últimos años, en parte gracias a los cambios en la posibilidad de recolectar y almacenar datos (Anderson, 2008) el interés en estas técnicas es cada vez más amplio (Tansley & Tolle, 2009; Lohr, 2012). El ámbito académico no es la excepción, la facilidad de capturar información de las trayectorias académicas o la interacción de los estudiantes con sistemas gestores de aprendizaje han permitido que estas técnicas puedan ser usadas por ejemplo en la personalización de contenidos en e-learning, analizar los patrones de uso o las actividades de los estudiantes, e incluso crear modelos de los tutores o estudiantes, (Romero & Ventura, 2010). Estas técnicas también se han usado para identificar o predecir el desempeño académico de los estudiantes, el cual puede ser entendido como aprobar o completar una unidad educativa, p.ej. un examen, un curso o un grado. (Kotsiantis, Pierrakeas & Pintelas, 2003; Superby, Vandamme & Meskens, 2006; Marquez-Vera, Romero & Ventura, 2010; López, León & González, 2015).

Bajo este supuesto, se trabajó con tres fuentes de datos: perfil integral, los datos de desempeño en la prueba de admisión, el puntaje básico de matrícula (PBM), los datos correspondientes a la primera matrícula del estudiante, incluyendo la variable de interés: si fue desvinculado o no de la Universidad durante este periodo.

Estos datos fueron analizados usando técnicas de minería de datos, las cuales fueron escogidas con base en su capacidad predic-tiva, interpretabilidad y la generación de

modelos menos propensos al sobreentrenamiento (*over-fitting*). Considerando estas características se escogieron inicialmente las siguientes técnicas: árboles de decisión, un clasificador Bayesiano y regresión logística.

Se usó la validación cruzada de 10 iteraciones (10-fold *cross-validation*) para probar el modelo y evitar el sobre entrenamiento. En esta, el conjunto de datos es dividido en 10 grupos, nueve son usados para ajustar el modelo y uno para probarlo. Este proceso se repite diez veces, en cada una de las iteraciones se deja uno de los grupos para prueba y se entrena con los nueve restantes.

Como criterios de evaluación se utilizó la sensitividad, o *recall*. Este es el porcentaje de registros correctamente clasificados en la clase de interés, en nuestro caso corresponde al porcentaje de estudiantes para quienes se predijo correctamente una desvinculación sobre el total de estudiantes que fueron desvinculados en la primera matrícula. Adicionalmente, una de las medidas de desempeño más comunes es el *accuracy* que corresponde al número de instancias correctamente clasificadas sobre el total de instancias; sin embargo, este criterio no funciona bien cuando el conjunto de datos no es balanceado, por ejemplo, si el 90% de los datos es de la clase A, un modelo que clasifique a todos los registros en esa clase tendría un *accuracy* de 90%. En estos casos se recomienda el uso de otras medidas de desempeño. El *balanced accuracy* tiene en cuenta esta característica y se calcula promediando el porcentaje de registros correctamente clasificados de cada una de las clases. En el ejemplo, dicha métrica sería 45%. Otra opción para manejar estas diferencias en los datos es otorgar pesos de acuerdo a los diferentes tipos de error, en este caso, es más grave no detectar a un estudiante en riesgo que detectar equivocadamente a alguien que no lo sea. Estos pesos son usados al momento de entrenar el modelo usando el algoritmo MetaCost (Domingos, 1999).

2.1 Población Objeto

El estudio tomó los datos de 2759 estudiantes pertenecientes a los 94 programas de pregrado de la Universidad que ingresaron en la cohorte del segundo semestre del 2013. Se escogió esta población porque de esta se disponía una mayor consistencia de datos en las siguientes fuentes de información:

1. Perfil integral: es una encuesta realizada por la Dirección Nacional de Bienestar de la Universidad con la que se busca determinar la condición general en la que ingresan los integrantes de la comunidad universitaria, hacer seguimiento a su permanencia y establecer las condiciones en las que egresa. La encuesta recoge información en cinco ámbitos: Académico-laboral, sociodemográfico, salud, dinámica personal y familiar, socioeconómico, intereses y prácticas deportivas, culturales, artísticas y comunitarias. Para el estudio se tomó información del componente académico-laboral, dinámica personal, sociodemográfico y socioeconómico, considerando que, de acuerdo a los planteamientos de Tinto (1993) y Giovagnoli (2002), pueden tener efectos cualitativos como cuantitativos en la definición de deserción estudiantil.

2. Prueba de admisión: es el instrumento con el que la Universidad evalúa el nivel académico de los aspirantes para que puedan ingresar como estudiantes. La Prueba de Admisión evalúa la comprensión de lenguaje y de los conceptos básicos requeridos para el estudio de las Matemáticas, las Ciencias Naturales y Físicas, las Ciencias Sociales y las Artes, en contextos y situaciones comunicativas. Los resultados obtenidos en la prueba en cada uno de los componentes evaluados, convierte los resultados de la prueba en un indicador aproximado del capital

académico con el que cuenta el estudiante al ingresar a la Universidad.

3. Puntaje Básico de Matrícula (PBM): "es un indicador sintético que pretende recoger la diferencia de las condiciones económicas de los estudiantes para el cobro de matrícula" (Pinto et al., 2007: 57) Este indicador cuenta con variables socio-demográficas y socioeconómicas que permiten conocer el capital económico con el que cuenta el estudiante para cursar sus estudios superiores.

3 Modelo propuesto

En una primera etapa se utilizaron los datos de estudiantes matriculados en el período académico 2013-II que tuvieran la información del perfil integral completa, 2759 registros, considerando que no todos los estudiantes diligenciaron la encuesta, porque es de carácter voluntario. La selección de atributos se realizó teniendo en cuenta los siguientes criterios.

- Análisis exploratorio del Perfil de Ingreso en el período académico 2013-03 en la Sede Bogotá

- Evaluación técnica (Gini)

- Factores referenciados en la literatura.

Inicialmente, el modelo fue construido a partir de:

Datos de Entrada: Puntaje examen de admisión en los cinco componentes, PBM, estrato, género, edad al ingreso e información sobre el programa.

A: Capacidad de adaptación, atención, perseverancia, manejo del tiempo y apoyo familiar, libertad en la elección de carrera y naturaleza del colegio.

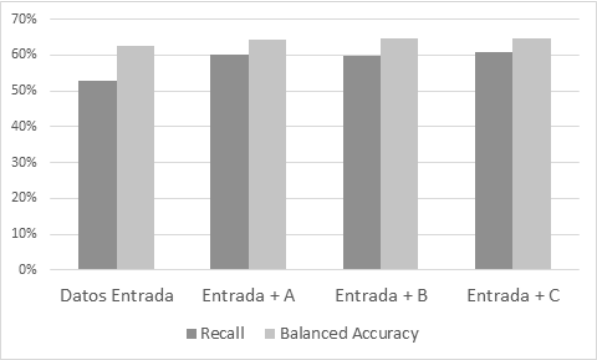
B: A - libertad en la elección de carrera y naturaleza del colegio.

C: B + percepción desempeño en matemáticas, percepción desempeño en lectoescritura y educación de la madre.

Además, se agregaron variables del perfil integral como la educación del padre, la propiedad de la vivienda, si debe mudarse a otro municipio al iniciar los estudios pero las medidas de desempeño no tuvieron grandes variaciones. Como puede verse en la Fig 2. el desempeño del modelo mejora, y, aunque el cambio no es notorio en la clasificación general (*balanced accuracy*) se logra un aumento de cerca de 7 puntos porcentuales en el *recall*.

En un segundo acercamiento, se favoreció la posibilidad de contar con más datos sobre la variedad de factores. Con base en esto, se usó únicamente la información de entrada, previa a la encuesta del perfil integral, de los estudiantes que ingresaron en los periodos académicos entre 2009-I y 2013-II. Dada la cantidad de datos, se crearon dos conjuntos, uno con las cohortes de 2009-I a 2013-II para entrenar el modelo (46782 registros) y otro para probarlo con la cohorte de 2013-II (4211 registros). Este es un escenario más realista pues el modelo se prueba con los "nuevos" estudiantes y se asemeja aún más a una tarea predictiva, p.ej. un sistema de detección temprana, donde se pueda detectar a los estudiantes en riesgo antes de su ingreso.

Figura 2. Medidas de desempeño



En este caso, el modelo se estrena con validación cruzada como fue descrito anteriormente y los resultados son muy cercanos a los mejores modelos de la primera

etapa, con un *recall* de 60% y *balanced accuracy* de 63%. Después, la predicción usando los datos de la nueva cohorte y en este caso las medidas de desempeño disminuyen a 57% y 61% respectivamente, pero siguen siendo mejores que el modelo usado con los mismos datos de entrada en la primera etapa. Puede verse entonces que incluir la información adicional acerca de la percepción del estudiante y aumentar el tamaño del conjunto de datos pueden tener efectos positivos en la calidad del modelo.

4 Conclusiones

Este trabajo presentó la implementación de técnicas de minería de datos con el fin de predecir los estudiantes desvinculados durante su primera matrícula. Las técnicas implementadas fueron regresión logística, árboles de decisión y Naïve Bayes, un clasificador Bayesiano; sin embargo, solo el último tuvo resultados aceptables. Una primera fase del trabajo consideró la cohorte de 2013-II en los 94 programas de pregrado de la Universidad Nacional de Colombia; en esta se incluyeron variables de desempeño académico, i.e. resultados en el examen de admisión, socioeconómicas y sociodemográficas, así como su percepción en el ámbito académico. Una segunda fase no tuvo en cuenta la percepción de los estudiantes y en su lugar trabajó con los datos académicos de las cohortes entre 2009-I y 2013-II.

El desempeño predictivo del modelo mejoró en dos situaciones, al aumentar el tamaño del conjunto de datos y también al incluir la información adicional acerca de la percepción del estudiante. Es importante resaltar que esta información adicional nos permite construir modelos con una cantidad menor de registros sin disminuir el desempeño, lo cual invita a pensar en formas que permitan capturar esta información a la espera de que el proceso de toma de decisiones se vea favorecido. Además de esto, nos brinda información más completa acerca de las características de los estudiantes y esto ofrece un interesante potencial ya que estas características podrán dar indicios de los programas de apoyo y de seguimiento que estos estudiantes puedan necesitar.

Se destacan los factores académicos como el programa del admitido y su desempeño en la prueba de admisión, principalmente en los componentes de ciencias y matemáticas; el programa académico al cual es admitido, el estado económico del estudiante, su edad y su percepción sobre sus habilidades. Estos resultados concuerdan con el trabajo de Gallón y Vasquez (2014) en la Universidad de Antioquia quienes destacan la edad, el apoyo económico y la adaptación académica y Rey y Diconca (2014) en la Universidad de la República quienes presentan el área de conocimientos y el capital cultural, que es representado por el nivel de educación de los padres.

Referencias

Anderson, C. (2008). The Petabyte age: because more isn't just more–more is different. Wired Magazine, (16.07), 106-120.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 155-164. ACM.

Gallón, S. & Vasquez, J. (2014) Aplicación de la Teoría de Clasificación al Problema del Abandono Estudiantil: Un estudio de Caso. Gestión Universitaria Integral de Abandono. IV CLABES. Cuarta Conferencia Latinoamericana sobre el Abandono en la Educación Superior. Libro de Actas. 43-52.

Giovagnoli, P. I. (2002). Determinantes de la deserción y graduación universitaria: Una aplicación utilizando modelos de duración. Documento de Trabajo 37. Universidad Nacional de la Plata.

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In Knowledge-

Based Intelligent Information and Engineering Systems (pp. 267-274). Springer Berlin Heidelberg.

Lohr, S. (2012). The age of big data. New York Times, 11.

Lopez, C., León, E., & Gonzalez, F. (2015). A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. IEEE Revista Iberoamericana de Tecnologías del Aprendizaje. 10(3), 1-7 doi: 10.1109/RITA.2015.2452632

Marquez-Vera, C., Romero, C., & Ventura, S. (2010). Predicting school failure using data mining. IEEE Revista Iberoamericana de Tecnologías del Aprendizaje. 8(1), 7-14. doi: 10.1109/RITA.2013.2244695

Pinto, M., Durán, D., Pérez, R., Reverón, C., & Rodríguez, A. (2007). Cuestión de supervivencia. Graduación, deserción y rezago en la Universidad Nacional de Colombia. Bogotá: Universidad Nacional de Colombia. Dirección Nacional de Bienestar Universitario.

Rey, R., Diconca, B. (2014). Factores Estructurales Asociados al Abandono en la Universidad de la República. Gestión Universitaria Integral de Abandono. IV CLABES. Cuarta Conferencia Latinoamericana sobre el Abandono en la Educación Superior. Libro de Actas. 53-62.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618.

Sistema para la Prevención de la Deserción Estudiantil,SPADIES. (15 de junio 2015). Recuperado de:http://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-propertyname-2895.html.

Superby, J. F., Vandamme, J. P., & Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. In Workshop on Educational Data Mining, 37-44.

Tansley, S., & Tolle, K. M. (Eds.). (2009). The fourth paradigm: data-intensive scientific discovery (Vol. 1). Redmond, WA: Microsoft Research.

Tinto, V. (1993). Reflexiones sobre el abandono de los estudios superiores. Perfiles Educativos, (62), 56-63.