Análisis de perfiles de rendimiento académico mediante técnicas de minería de datos y análisis de datos multivariados

Maria Paula DIESER[(1)], María Cristina MARTÍN[(1)(2)], Lorena Veronica CAVERO[(1)], Sofía FUNKNER[(1)], Laura WAGNER[(1)]

[(1)] Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa
[(2)] Departamento de Matemática, Universidad Nacional del Sur

{pauladieser, maritamartin}@exactas.unlpam.edu.ar

RESUMEN

La deserción estudiantil, especialmente en los primeros años de la carrera, es una preocupación presente y constante en todas las Instituciones de Nivel Superior. Para un tratamiento efectivo y eficaz del problema, resultan indispensables la detección temprana de estudiantes en situación de riesgo en términos de abandono o retraso en el alcance del grado, y el diseño e implementación de un plan de acción consecuente.

La Facultad de Ciencias Exactas y Naturales de la Universidad Nacional de La Pampa no es ajena a esta realidad. En consecuencia, la línea de investigación aquí presentada, propone estudiar y aplicar distintos métodos que ofrece la Minería de Datos y el Análisis de Datos Multivariados sobre los datos registrados en el sistema de gestión de información estudiantil de la Institución con el propósito de caracterizar la trayectoria académica de los estudiantes, y detectar patrones compatibles con situaciones de dificultades en el aprendizaje, que puedan derivar en el abandono de los estudios.

Palabras clave: minería de datos educativos, análisis de datos multivariados, rendimiento académico, deserción universitaria.

CONTEXTO

La línea de investigación que aquí se describe, se instala dentro de un Proyecto de Investigación más amplio, acreditado y financiado por la Facultad de Ciencias Exactas y Naturales (FCEyN) de la Universidad Nacional de La Pampa (UNLPam). Dicho Proyecto se deriva de tareas de investigación desarrolladas en la Institución durante el periodo 2014 – 2017, y vinculadas con el estudio y aplicación de métodos multivariados discriminantes y de clasificación, con el propósito de establecer similitudes y diferencias, y analizar las estimaciones que se obtienen con ellos al aplicarlos efectivamente en el Análisis de Datos Multivariados.

De las investigaciones realizadas, surge el campo de la educación como un terreno propicio para la aplicación de técnicas de Minería de Datos (MD), que pueden complementarse con otras propias del Análisis de Datos Multivariados (ADM). Además, tales métodos pueden nutrirse con elementos de la Teoría de Respuesta al Ítem (TRI) y el Análisis de Supervivencia (AS) para el análisis de las respuestas en cuestionarios y del tiempo requerido para la aprobación de asignaturas o la graduación, respectivamente.

En particular, la línea de investigación que aquí se presenta, iniciada en 2018, tiene por objetivo general estudiar y aplicar distintos métodos que ofrece el ADM y la MD sobre los datos registrados en SIU Guaraní de la FCEyN (UNLPam) con el propósito de caracterizar la trayectoria académica de los estudiantes, y detectar patrones compatibles con situaciones de dificultades en el aprendizaje, que puedan resultar en abandono de los estudios.

1. INTRODUCCIÓN

La comunidad universitaria en su conjunto se plantea y propone la mejora continua de la calidad de los procesos educativos que se desarrollan en sus instituciones, y de los servicios que ofrecen. La FCEyN (UNLPam) no es ajena a esta realidad. El equipo de

gestión, cuerpo docente y agrupaciones estudiantiles, a través de la Comisión *ad hoc* de Ingreso y Permanencia (CIP), han diagnosticado altos niveles de deserción y desgranamiento en los primeros años de estudio. No obstante, los diagnósticos realizados carecen de la sistematización necesaria que permita revelar a tiempo el abandono de estudiantes en diferentes tramos de las carreras elegidas.

Al mismo tiempo, en el proceso de inscripción a las carreras de grado de la FCEyN (UNLPam), y en el desarrollo de las actividades del Programa de Ambientación a la Vida Universitaria (PAVU) de la Institución, se recolectan múltiples datos aportados por los aspirantes a través de los sistemas de gestión de información que luego son enriquecidos con datos relativos a la historicidad académica de los estudiantes. Éstos constituyen una importante fuente de información, en la medida que se extraiga conocimiento para el análisis de la realidad de los estudiantes y los contextos en los que ellos aprenden, y para el diseño de eventuales planes de acción.

La MD, en combinación con el ADM, reúnen un conjunto de técnicas capaces de modelizar y resumir dicha información, facilitando su comprensión y ayudando a la toma de decisiones en situaciones futuras (Cabena et al. 1998; Hernández Orallo et al., 2004). En particular, la Minería de Datos Educativos (MDE) se presenta como un área iluminada por diferentes disciplinas, relativamente reciente y de crecimiento notable, que se ocupa del desarrollo, la investigación y la aplicación de métodos computacionales para detectar patrones en grandes conjuntos de datos educativos que, de otro modo, serían difíciles o imposibles de analizar debido a su volumen (Romero & Ventura, 2010).

Revisiones de investigaciones realizadas en MDE dan cuenta de los objetivos perseguidos y las diversas aplicaciones posibles en el área (Romero & Ventura, 2007, 2010; Baker & Yacef, 2009). Romero & Ventura (2010), en base a estas revisiones, elaboran una taxonomía de las áreas de aplicación de MDE, entre las que se menciona la predicción del desempeño académico de los estudiantes.

Sin embargo, el estudio del rendimiento académico y del abandono escolar no es de interés reciente, y siempre ha estado relacionado con factores sociales, económicos y psicológicos. Varios estudios han abordado estos temas usando distintas metodologías: análisis discriminante, reglas de asociación, modelos de regresión logística, ANOVA, árboles de decisión, redes neuronales, redes bayesianas, entre otros (Streeter & Franklin, 1991; Ma et al., 2000; Wayman, 2001; Pursley, 2002; Minaei-Bidgoli et al., 2003; Kotsiantis et al., 2004; Pardos et al., 2006; Cortez & Silva, 2008; Márquez Vera et al., 2012).

La línea de investigación que aquí se describe pretende realizar un aporte desde el área sobre la realidad y contexto de la FCEyN (UNLPam), proporcionando modelos que permitan caracterizar la trayectoria académica de los estudiantes, y detectar patrones compatibles con situaciones de dificultades en el aprendizaje y abandono. Estos modelos podrían ser de utilidad para implementar políticas de retención adecuadas.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Como se mencionó anteriormente, la línea de investigación aquí presentada se enmarca en un Proyecto más amplio y tiene por objetivo general estudiar distintas técnicas de la MD y el ADM utilizadas en el campo educativo para la obtención de modelos de estudiantes que permitan identificar situaciones de riesgo de deserción o abandono.

Las técnicas estudiadas serán aplicadas sobre los datos registrados en el sistema de gestión de información estudiantil (SIU Guaraní) de la FCEyN (UNLPam) mediante rutinas desarrolladas en el lenguaje de programación R y, eventualmente, otros programas existentes para el procesamiento de datos (*e.g.* RapidMiner, STATGRAPHICS, STATA, SPSS, S-PLUS, NTSys, STATISTICA).

Así, los resultados obtenidos serán evaluados y comparados de manera que los mejores modelos sean utilizados en la identificación

temprana de estudiantes en riesgo, y el establecimiento de una política de apoyo académico adecuada para atender la situación y, eventualmente, disminuir los índices de fracaso y abandono.

3. RESULTADOS OBTENIDOS/ESPERADOS

A lo largo del primer año se han desarrollado las tareas que se enumeran a continuación:

- Revisión sistemática de bibliografía referida a experiencias desarrolladas en el ámbito de la Educación Superior que utilicen la MDE y el ADM para identificar modelos que describen la trayectoria académica de estudiantes y patrones de deserción o abandono, poniendo especial atención a las técnicas y *software* utilizados, los atributos considerados, y aquéllos que se vinculan de manera significativa con el rendimiento académico.
- Estudio pormenorizado de las técnicas empleadas en las investigaciones empíricas revisadas, y otras de eventual utilidad, apoyado por el desarrollo de prácticas y aplicaciones sencillas mediante el lenguaje R y *software* estadístico de utilidad.
- Análisis de datos censales disponibles en SIU Guaraní (historia académica, atributos personales, y de índole social y económica) correspondientes a 9187 registros de estudiantes que han asentado su ingreso a la Institución a partir del año 2001 y hasta 2018.

A partir de estas tareas previas, y a fin de alcanzar los objetivos propuestos, se espera en los próximos años llevar adelante las siguientes acciones, aún pendientes:

- Aplicación de las técnicas estudiadas, y otras que pudieran emerger como potencialmente útiles, sobre los datos provenientes de SIU Guaraní de la FCEyN (UNLPam), previo desarrollo de técnicas de preprocesamiento (limpieza, selección de variables, y la transformación o combinación de éstas) que permitan obtener una vista minable de los datos recopilados.
- Evaluación y comparación de los patrones y modelos resultantes a partir de un análisis e interpretación del conocimiento obtenido. Esto permitirá seleccionar los modelos más expresivos, para finalmente elaborar conclusiones pertinentes y comunicar los resultados alcanzados.

Se espera así, en un plazo total no superior a los cinco años, contribuir a la identificación temprana de estudiantes en riesgo, y el establecimiento de estrategias académicas adecuadas para atender la situación y, eventualmente, disminuir los índices de fracaso y abandono.

4. FORMACIÓN DE RECURSOS HUMANOS

En la línea de investigación presentada, bajo la dirección de la Dra. Martín, trabajan tres docentes investigadoras cuya formación de base corresponde al campo de la matemática o la educación matemática.

- Dos de ellas han finalizado el cursado de la Maestría en Tecnología Informática Aplicada en Educación de la Facultad de Informática de la Universidad Nacional de La Plata y se encuentran en proceso de elaboración del proyecto de tesis. Una de ellas proyecta trabajar en temas de autorregulación del aprendizaje y su impacto en el rendimiento académico de estudiantes universitarios. Los resultados derivados del proceso de investigación correspondiente podrían abrevar los objetivos propuestos para la línea de investigación descripta, y ampliar el rango de datos y atributos estudiados originalmente.
- La tercera ha comenzado sus estudios de Doctorado en Estadística en la Universidad Nacional de Rosario y proyecta realizar su trabajo de tesis doctoral en el área del ADM. Los conocimientos alcanzados podrían nutrir los procesos desarrollados al interior de este Proyecto y, en particular, los propios de la línea de investigación que aquí se presenta.

El equipo de trabajo cuenta también con la participación de una estudiante avanzada de Licenciatura en Matemática que ha orientado su formación específica en temas de estadística aplicada. Como tal, ha obtenido una Beca de Iniciación a la Investigación otorgada por la UNLPam, para desarrollar el proyecto "Los Árboles de Decisión aplicados al análisis de Datos Educativos" bajo la dirección de la Lic. Dieser. Se espera que los resultados alcanzados durante el periodo de Beca permitan a la postulante especializarse en una de las técnicas abordadas en esta línea de investigación y eventualmente, iniciar estudios de postgrado en estos temas u otros vinculados.

En cualquier caso, se espera originar y organizar un plantel humano (docentes, investigadores, y estudiantes) que aporte información de utilidad para generar políticas adecuadas en relación al ingreso y permanencia de estudiantes en la FCEyN (UNLPam). Asimismo, se considera la posibilidad de desarrollar otros temas de interés relacionados con las investigaciones que se realizan en el marco de esta línea y del Proyecto en general, que pudieran surgir como consecuencia de replanteos o de consultas atendidas en el asesoramiento estadístico a otros investigadores o instituciones del medio.

5. BIBLIOGRAFÍA

Baker, R. S. J. D. & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1(1):3–16.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: from concept to implementation*. New Jersey: Prentice Hall.

Cortez, P. & Silva, A. (2008). Using data mining to predict secondary school student performance. En Brito, A. and Teixeira, J. (Eds.), *Proceedings of 5th Future Business Technology Conference*, pp. 5–12, Porto, Portugal. EUROSIS.

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la Minería de Datos*. Madrid: Pearson Prentice Hall.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting student's performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18(5):411–426.

Ma, Y., Liu, B., Wong, C. K., Yu, P. S., & Lee, S. M. (2000). Targeting the right students using data mining. En *Proceedings of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 457–464, Boston, USA.

Márquez Vera, C., Romero Morales, C., & Ventura Soto, S. (2012). Predicción del Fracaso Escolar Mediante Técnicas de Minería de Datos. *IEEE-RITA*, 7(3):109–117.

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. En *Proceedings of 33rd Annual Frontiers in Education, FIE 2003*, pp. 13–18, Colorado, USA.

Pardos, Z. A., Heffernan, N. T., Anderson, B., and Heffernan, C. L. (2006). Using fine-grained skill models to fit student performance with bayesian networks. En *Proceedings of the Workshop in Educational Data Mining held at the 8th International Conference on Intelligent Tutoring Systems*, Taiwan.

Pursley, M. (2002). *Changes in Personal Characteristics of Mexican-American High School Graduates and Dropouts During the Transition from Junior High to High School*. Texas Tech University.

Romero, C. & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.*, 33(1):135–146.

Con formato: Inglés (Estados Unidos)

Con formato: Inglés (Estados Unidos)

Con formato: Inglés (Estados Unidos)

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

Streeter, C. L. & Franklin, C. (1991). Psychological and family differences between middle class and low income dropouts: A discriminant analysis. *The High School Journal*, 74(4):211–219.

Wayman, J. C. (2001). Factors influencing GED and diploma attainment of high school dropouts. *Education Policy Analysis Archives*, 9(4):1–19.