APLICACIÓN DE TÉCNICAS ESTADÍSTICAS Y DE MINERÍA DE DATOS PARA EL ANÁLISIS DE PERFILES DE RENDIMIENTO ACADÉMICO: EL CASO DE LA FACULTAD DE CIENCIAS EXACTAS Y NATURALES DE LA UNIVERSIDAD NACIONAL DE LA PAMPA

Lorena Veronica CAVERO[(1)], Maria Paula DIESER[(1)], María Cristina MARTÍN[(1)(2)], Sofia FUNKNER[(1)], Janina ROLDAN[(1)], Erica SCHLAPS[(3)], Diamela TITIONIK[(1)], Laura WAGNER[(1)]

[(1)] Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa
[(2)] Departamento de Matemática, Universidad Nacional del Sur
[(3)] Instituto de Ciencias Polares, Ambiente y Recursos Naturales, Universidad Nacional de Tierra del Fuego, Antártida e Islas del Atlántico Sur

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

RESUMEN

En el proceso de inscripción a las carreras de grado de la Facultad de Ciencias Exactas y Naturales de la Universidad Nacional de La Pampa, y a lo largo del recorrido académico que hacen los estudiantes por la institución, se recolectan múltiples datos a través de los sistemas de gestión. Éstos constituyen una importante fuente de información, en tanto se extraiga conocimiento útil 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. En particular, interesa a la comunidad institucional la detección temprana de estudiantes en situación de riesgo en términos de deserción o retraso en el alcance del grado.

La línea de investigación presentada, propone estudiar y aplicar distintos métodos que ofrece la Minería de Datos, el Análisis de Datos Multivariados, la Teoría de Respuesta al Ítem y el Análisis de Supervivencia, 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 abandono de los estudios.

Palabras clave: minería de datos, análisis de datos multivariados, teoría de respuesta al ítem, análisis de supervivencia, deserción universitaria.

CONTEXTO

Durante el periodo 2014 - 2017 se realizaron tareas de investigación, en el ámbito de la Facultad de Ciencias Exactas y Naturales (FCEyN) de la Universidad Nacional de La Pampa (UNLPam), vinculadas con el estudio y aplicación de métodos multivariados discriminantes y de clasificación (algunos que podrían entenderse como clásicos y de una esencia más estadística, y otros propios de la minería de datos) 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 (ADM).

De las investigaciones realizadas, surge el campo de la educación como un terreno propicio para las aplicaciones de Minería de Datos (MD), dada la multiplicidad de fuentes de datos y los diversos grupos de interés implicados. Asimismo, el área educativa ofrece la posibilidad de aplicar 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 consecuencia, se inicia en 2018 un nuevo Proyecto, acreditado y financiado por la Institución mencionada, cuyo objetivo general es estudiar y aplicar distintos métodos que ofrece la MD, el ADM, la TRI y el AS, sobre los datos registrados en SIU Guarani 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 derivar en abandono de los estudios universitarios.

1. INTRODUCCIÓN

En la actualidad, la mayoría de los procesos (industriales, académicos, negocios, servicios, entre otros) cuentan con información histórica almacenada. El avance de la tecnología ha permitido generar volúmenes de datos cada vez más grandes y difíciles de analizar y comprender. Distintas áreas han tratado de dar soluciones a este problema. La MD, en combinación con el ADM, reúnen un conjunto de técnicas capaces de modelizar y resumir la información, facilitando su comprensión y ayudando a la toma de decisiones en situaciones futuras (Cabena et al. 1998; Hernández Orallo et al., 2004). El área educativa no escapa a esta realidad. En general, los establecimientos educativos disponen de información sumamente detallada de cada alumno proveniente de múltiples fuentes (*e.g.* bases de datos, aplicaciones web, entornos virtuales de enseñanza y aprendizaje) pero carecen de modelos que les permitan describir de forma objetiva a sus estudiantes. Caracterizar a los estudiantes de una institución académica aporta información no trivial y de utilidad para la toma de decisiones.

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.

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 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.

Este tipo de estudios en el campo de la educación corresponde a aplicaciones de una rama particular de la MD conocida como Minería de Datos Educativos (MDE). Este nuevo área de investigación interdisciplinaria se ocupa del desarrollo de métodos para explorar los datos que se dan en el ámbito educativo, así como de la utilización de estos métodos para entender mejor a los estudiantes y los contextos en que ellos aprenden (Romero & Ventura, 2010). Romero et al. (2010) definen la MDE como el desarrollo, investigación y 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.

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 de 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 y de imputación múltiple, 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).

Asimismo, el área educativa ofrece la posibilidad de aplicar elementos de la TRI y el AS. En particular, la TRI ofrece estimaciones del rasgo latente de individuos medidos mediante un test o cuestionario. Su utilidad en el campo educativo radica en determinar si un estudiante consigue responder correctamente a cada una de las preguntas que componen el cuestionario y en atender al puntaje bruto obtenido en la prueba (Bartholomew & Knot, 1980; Bartholomew et al., 2008; Burga León, 2005; Debera & Nalbarte, 2006; Hidalgo Flores, 2007; Pardo Adames, 2001). El AS, por su parte, aporta técnicas que permiten extraer conclusiones del tiempo requerido para la aprobación de espacios curriculares o la graduación, así como su relación con predictores sociodemográficos y de aptitud académica, entre otros (Rojas Torres & Alfaro Rojas, 2014; Gallardo Allen et al., 2016).

En suma, la MDE mediante técnicas de MD, ADM, TRI, y AS es un área de investigación relativamente reciente y de crecimiento notable. 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 desprende de un Proyecto anterior que permitió investigar técnicas de discriminación y clasificación multivariadas, con el propósito de establecer similitudes o diferencias, y analizar la eficiencia de las mismas al aplicarlas en el análisis de datos multivariados.

Habiendo identificado el campo de la educación como un terreno propicio para las aplicaciones de muchas de las técnicas estudiadas, en esta nueva línea se pretende estudiar y aplicar distintos métodos que ofrecen la MD y el ADM, la TRI y el AS, sobre los datos registrados en el sistema de gestión de información estudiantil (SIU Guarani) 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 derivar en abandono de los estudios universitarios. Los resultados obtenidos serán evaluados y comparados de manera que los mejores modelos resultantes podrían ser de utilidad 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

Debido a que la presente línea de investigación recién se inicia, no se cuenta a la fecha con resultados propios.

Hasta el momento se ha realizado una revisión sistemática de bibliografía referida a experiencias que utilicen la MDE para identificar modelos que describen la trayectoria académica de estudiantes y patrones de deserción o abandono, poniendo especial atención a las técnicas utilizadas vinculadas con el ADM, la MD, la TRI, y el AS. En particular, interesan aquellas experiencias en el ámbito de la Educación Superior de la República Argentina.

Habiendo identificado las técnicas empleadas en los estudios empíricos revisados, y considerando otras que pudieran resultar de utilidad, se espera comenzar con su aplicación sobre los datos provenientes de SIU Guarani de la FCEyN (UNLPam), previo desarrollo de técnicas de preprocesamiento (limpieza, transformación, selección de variables, y la transformación o combinación de éstas) que permitan obtener una vista minable de los datos recopilados.

Aplicadas las técnicas seleccionadas, se evaluarán y compararán 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 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, y la co-dirección de la Lic. Dieser, trabajan cuatro docentes/auxiliares investigadoras (tres perteneciente a la FCEyN (UNLPam) y una de la Universidad Nacional de Tierra del Fuego, Antártida e Islas del Atlántico Sur), todas con formación de base matemática. Una de ellas ha 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. Las restantes han comenzado sus estudios de Doctorado en Estadística en la Universidad Nacional de Rosario, y una de ellas proyecta realizar su trabajo de Tesis Doctoral en AS, línea que, como ya se manifestara, se plantea aplicar para el estudio de la permanencia de los estudiantes universitarios.

El equipo de trabajo cuenta también con la participación de una Becaria, auxiliar docente en la FCEyN (UNLPam) quien lleva adelante su proyecto de tesis, bajo la dirección de la Dra. Martín, "La Teoría de Respuesta al Ítem aplicada a prueba diagnóstico de ingreso universitario", a fin de obtener el grado de master en la Maestría en Estadística Aplicada de la Universidad Nacional de Córdoba. Finalmente, como asistente de investigación, se incorpora una estudiante de Licenciatura en Matemática que ha orientado su formación específica en temas de estadística aplicada. Se espera que pueda aplicar aquí los aprendizajes apropiados, adquirir nuevos conocimientos, y eventualmente, iniciar estudios de postgrado en temas vinculados con los de este Proyecto.

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.

Bartholomew, D.J. & Knott, M. (1980). Latent Variables Modelos and Factor Analysis. Kendall´s Library of Statistics, 1° Edition.

Bartholomew D.J.; Steele, F., Moustaki, I. & Galbraith, J.I. (2008). Analysis of multivariate social science data. 2° Edition. Boca Ratón, EEUU: Taylor & Francis Group.

Burga León A. (2005). Evaluación del rendimiento académico. Introducción a la Teoría de Respuesta al Ítem. Ministerio de Educación, Lima. Perú.

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.

Debera, L. & Nalbarte, L. (2006). *Pruebas diagnósticas: una aplicación a la teoría de respuesta al ítem, aproximación clásica y bayesiana*. Instituto de Estadística. F.C.E. y Administración, Universidad de la República.

Gallardo Allen, E, Molina Delgado, M. & Cordero Cantillo, R. (2016). Aplicación del Análisis de Sobrevivencia al Estudio del Tiempo Requerido para Graduarse en Educación Superior: El Caso de la

Universidad de Costa Rica. *Páginas de Educación*, 9(1):61–87.

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

Hidalgo Flores, R. (2007). *Teoría de respuesta al ítem: una aplicación educativa*. Facultad de Ingeniería, Universidad Autónoma de Querétaro, México.

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.

Pardo Adames, C. (2001). El modelo de Rasch: Una alternativa para la evaluación educativa en Colombia. Facultad de Psicología. Universidad Católica de Colombia.

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.

Rojas Torres, L. & Alfaro Rojas, L. (2014). Análisis de sobrevivencia para la estimación del tiempo adicional como adecuación para la aplicación de una prueba estandarizada. PEL: Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana. 51(1), 135-155.

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

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.

Romero, C., Ventura, S., Pechenizky, M., & Baker, R. (2010). *Handbook of Educational Data Mining*. Chapman and Hall CRC Press, Taylor & Francis Group, Boca Raton.

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.