Evaluación de Técnicas de Clasificación para Predecir el Rendimiento Académico de Ingresantes a la Universidad en Temas de Matemática

Maria Paula DIESER[(1)], Lorena Veronica CAVERO[(1)], María Cristina MARTÍN[(1)(2)], Erica SCHLAPS[(1)], 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

{pauladieser, cavero, 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 en el desarrollo de las actividades del Programa de Ambientación a la Vida Universitaria de la Institución, se recolectan múltiples datos aportados por los aspirantes a través de los sistemas de gestión. Éstos constituyen una importante fuente de información, en tanto 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. Es una realidad la constante preocupación de la comunidad institucional por los elevados índices de deserción, o retrasos en alcanzar su título de grado, por dificultades en asignaturas vinculadas con la matemática.

La línea de investigación presentada propone procesar los datos recolectados a través de los sistemas de gestión durante el ingreso, y resultados del seguimiento de la actividad académica en asignaturas de matemática, para obtener posibles patrones entre los estudiantes que alcancen idénticos logros. Los modelos resultantes permitirán predecir el rendimiento académico en el área y determinar factores que lo afectan para implementar políticas de retención adecuadas.

Palabras clave: clasificación, predicción, minería de datos, rendimiento académico

Contexto

Desde 2014 se vienen realizando tareas de investigación, en el ámbito de la Facultad de Ciencias Exactas y Naturales (FCEyN) de la Universidad Nacional de La Pampa (UNLPam), relacionadas con el estudio y aplicaciónde métodos multivariados de discriminación 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. El Proyecto, acreditado y financiado por la Institución mencionada, ha contado también con la participación de estudiantes de postgrado de la Universidad Nacional de Asunción (UNA).

Entre los métodos estudiados en el marco del Proyecto, se encuentran algunos que podrían entenderse como clásicos y de una esencia más estadística (discriminación debida a Fisher), y otros propios del *Data Mining* (Árboles y Reglas de Clasificación, Redes Neuronales, y el Análisis de *Clusters*). Se ha desarrollado la teoría sobre estas técnicas, y aplicado a diferentes conjuntos de datos a fin de analizar su sensibilidad y fiabilidad, realizando prácticas con el lenguaje de programación R.

De las investigaciones realizadas, surge el campo de la educación como un terreno propicio para las aplicaciones de *Data Mining*, 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 para el análisis de las respuestas en cuestionarios, y el Análisis de Supervivencia, para extraer conclusiones del tiempo requerido para la aprobación de espacios curriculares o la graduación.

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 de la 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, en muchos casos asociados a losbajos rendimientos en asignaturas vinculadas con la matemática. 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.

Entre las políticas de gestión impulsadas por la CIP, se organiza en cada ciclo lectivo, previo al inicio de las cursadas regulares, una seriede acciones en el marco del Programa de Ambientación a la Vida Universitaria (PAVU) que incluye charlas, talleres y actividades recreativas destinadas a los aspirantes. Entre éstas, desde 2015, se desarrolla el Taller "Introducción a la Matemática" cuyo propósito es recuperar los conocimientos de matemática elemental que poseen los ingresantes y que son requeridos para el cursado de las asignaturas del área, contempladas en la oferta académica de grado de la FCEyN (UNLPam).

El Taller está a cargo de docentes y auxiliares docentes del Departamento de Matemática de la institución y cuenta con la colaboración de estudiantes avanzados de Profesorado y Licenciatura en Matemática en las tareas de tutorías. Es de carácter semipresencial, durante las tres semanas previas al inicio del primer período de clases. Los encuentros presenciales se distribuyen en 8 encuentros de 2 horas reloj cada uno, y están acompañados por actividades diversas implementadas sobre el curso *online* del Taller, desarrollado en el entorno virtual de enseñanza y aprendizaje *Moodle*. Este curso se estructuró en tres temas correspondientes a los tres bloques temáticos considerados para el tratamiento de los contenidos (Números, Álgebra, y Funciones). Se pusieron a disposición de los estudiantes, los materiales diseñados *ad hoc* (apuntes teóricos y trabajos prácticos), así como diferentes foros destinados a la comunicación de las novedades del Taller, el establecimiento de lazos sociales, y la evacuación de dudas y consultas. A partir de 2016, finalizado el Taller, se propuso a los estudiantes completar una autoevaluación de los contenidos trabajados y un cuestionario diseñado con preguntas cerradas vinculadas con diversos aspectos de índole demográfica, social, emocional, y escolaridad previa que se supone pueden afectar el rendimiento del estudiante. Estos datos, junto con los referidos a la asistencia al Taller y la participación en las actividades del curso *online* asociado, son considerados para analizar la influencia de las variables involucradas en el rendimiento académico de los estudiantes en asignaturasde matemática cursadas en el primer año de sus carreras. La finalidad es obtener información útil para 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.

Este tipo de estudios en el campo de la educación corresponde a aplicaciones de una rama particular del *Data Mining*(DM) conocida como Minería de Datos Educativos (EDM, por sus siglas en inglés). Este nuevo espacio de investigación interdisciplinario se ocupa del desarrollo y utilización de métodos para explorar los datos que se dan en el ámbito educativo, así como también para entender mejor a los estudiantes y los contextos en que ellos aprenden (Romero & Ventura, 2010). Romero et al. (2010) definen la EDM 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 EDM dan cuenta de los objetivos perseguidos y las diversas aplicaciones posibles en el área (Romero & Ventura, 2007, 2010; Baker & Yacef, 2009). En particular, Romero & Ventura (2010) elaboran una taxonomía de las áreas de aplicación de EDM, entre las que se menciona la predicción del desempeño de estudiantes.

Sin embargo, el estudio del rendimiento académico de los estudiantes y el 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, reglasde asociación, modelos de regresión logística y de imputación múltiple, análisis de la varianza, 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).

Por otra parte, la Teoría de Respuesta al Ítem (TRI) y el Análisis de Supervivencia (AS) son áreas de investigación estadística que podrían ofrecer técnicas adecuadas para el análisis de datos educativos. En particular, la TRI ofrece estimaciones del rasgo latente de individuos medidos mediante un test o cuestionario (Hidalgo Flores, 2007). 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 (Debera & Nalbarte, 2006). Por su parte, el AS permite modelizar el tiempo que se tarda en que ocurra un determinado suceso y su dependencia con otras posibles variables explicativas. En el ámbito de la educación, aporta técnicas apropiadas para analizar el tiempo requerido para graduarse o alcanzar determinado objetivo, así como su relación con predictores sociodemográficos y de aptitud académica, entre otros (Gallardo Allen et al., 2016).

2. Líneas de Investigación y Desarrollo

La línea de investigación aquí presentada surge de un Proyecto más amplio cuyo objetivo general es investigar técnicas de discriminación y clasificación multivariadas, con el propósito de establecer similitudes o diferencias y analizar la eficiencia de las mismas al aplicarlas, efectivamente, en el análisis de datos multivariados. En esta línea se pretende aplicar técnicas de DM para predecir el rendimiento académico de estudiantes ingresantes a la FCEyN (UNLPam) en asignaturas vinculadas con la matemática, empleando diversos métodos de clasificación (reglas y árboles de calificación, y redes neuronales), en combinación con otros propios de la TRI y el AS. Las técnicas seleccionadas serán aplicadas sobre una base de datos construida a partir de distintas fuentes: cuestionarios con preguntas cerradas vinculadas con diversos aspectos de índole demográfica, social, emocional, y escolaridad previa que se supone pueden afectar el rendimiento del estudiante; resultados obtenidos del Taller "Introducción a la Matemática" desarrollado durante 2016 (asistencia, participación, y autoevaluación); e informes finales incluyendo las calificaciones obtenidas por los estudiantes en las asignaturas de matemática cursadas en el primer año de las carreras respectivas. Los resultados obtenidos serán comparados, y 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 y Esperados

Hasta el momento se han llevado a cabo la fase de integración y recopilación de datos (determinando las fuentes de información descriptas anteriormente consideradas de utilidad para conformar una base de datos unificada); y la fase de limpieza y transformación como parte del

preprocesamiento de los datos (detectando la existencia de ciertos "problemas" en los datos, analizando y decidiendo formas adecuadas para su tratamiento). Como resultado de estos procesos se ha obtenido una vista minable de los datos recopilados.
Actualmente se están realizando las tareas vinculadas con la selección de variables relevantes para el objetivo de estudio, y la transformación o combinación de estas últimas. Esto permitirá reducir la dimensionalidad del problema de manera adecuada para simplificar el trabajo en las fases posteriores, a saber, (a) selección del método que produzca los patrones y modelos más expresivos; (b) evaluación de los patrones obtenidos a partir de un análisis e interpretación del conocimiento obtenido; y (c) comparación de los modelos obtenidos con los que pudieran surgir de la aplicación de otras técnicas de clasificación clásicas.
Se espera que los resultados alcanzados contribuyan a la identificación temprana de estudiantes en riesgo, y al establecimiento de estrategias académicas adecuadas para atender la situación y, eventualmente, disminuir los índices de fracaso y abandono.
Esta línea de investigación podría dar origen a un nuevo Proyecto más amplio en el que se consideren los datos registrados en el sistema de gestión de información estudiantil (SIU Guarani) de la FCEyN (UNLPam) a fin de vincular la información referida al rendimiento académico de los estudiantes con aquellos de índole socioeconómica, familiar, escolaridad previa, entre otros.

4. Formación de Recursos Humanos

En el área del Proyecto de Investigación, bajo la Dirección de la Dra. Martín, se han formado dos de los integrantes como graduados de la Maestría en Estadística y Metodología de la Investigación Científica Básica y Aplicada de la FCEyN (UNA) y, otros dos como egresadas de la Licenciatura en Matemática de la FCEyN (UNLPam), en 2014 y 2016, respectivamente. En la línea aquí presentada (DM), de un total de once integrantes, trabajan dos con formación de base matemática, computacional y pedagógica, quienes finalizaron el cursado de la Maestría en Tecnología Informática Aplicada en Educación de la Facultad de Informática (UNLP) y se encuentran en proceso de elaboración del proyecto de tesis. Además, tres de las integrantes 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. Respecto de la otra línea de investigación (TRI) a estudiar, a la brevedad, se solicitará al Consejo Directivo de la FCEyN (UNLPam) la incorporación de una nueva integrante del proyecto, quien actualmente se desempeña como auxiliar docente en la Institución y ha obtenido una beca del Programa Becas de Investigación y Postgrado / Subprograma Becas de Postgrado para iniciar Doctorados y Maestría de la UNLPam, por el período 2017-2018. El plan propuesto, también bajo la dirección de la Dra. Martín, es "La Teoría de Respuesta al Ítem aplicada a prueba diagnóstico de ingreso universitario", y busca la obtención del grado de *master* en la Maestría en Estadística Aplicada de la Universidad Nacional de Córdoba. Finalmente, cabe señalar que, se prevé sumar estudiantes de grado interesados en esta línea de investigación.

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.

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.

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.

Pardos, Z. A., Heffernan, N. T., Anderson, B., and Heffernan, C. L. (2006). Using fine-grained skill models to fit studentperformance 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.

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.