7

Modelo de Predicción y Prevención del abandono estudiantil en enseñanzas a distancia basado en minado web

David Lizcano
Universidad a Distancia de Madrid, UDIMA, España

Introducción

La sociedad actual demanda de forma constante una reflexión profunda sobre las metodologías pedagógicas utilizadas en todos los niveles educativos, para dar respuesta a la diversidad y flexibilidad necesaria para que cualquier ciudadano pueda acceder, en tiempo, localización y forma, a una educación de calidad. Una respuesta a estas necesidades son los sistemas de aprendizaje a distancia (*e-learning*), que, aunque tienen grandes beneficios, también plantean grandes desafíos. Uno de ellos es cómo hacer un buen trabajo de tutoría con estudiantes sin mantener un contacto directo con ellos, para evitar que caigan en el desánimo y el abandono en planes educativos a distancia que supongan auténticos retos por su complejidad y extensión. Existen datos alarmantes de abandono en todas las instituciones universitarias que emplean una metodología de enseñanza e-learning (según estudios de 2018, la tasa de abandono supera en un 20% a la de las mismas enseñanzas en su modalidad presencial), datos que se agravan aún más en enseñanzas superiores técnicas e ingenierías, lo que a todas luces indica que se trata de un problema abierto aún por resolver (Hershkovitz y Nachmias 2008; Kotsiantis *et al.,* 2003; Lykourentzou *et al.,* 2009; Madhyastha y Hunt 2009).

El objetivo de esta propuesta consiste en lograr un mejor seguimiento de los estudiantes gracias a las herramientas pedagógicas y de seguimiento empleadas, y prevenir a tiempo casos de posible abandono. Las metodologías on-line, así como las plataformas educativas empleadas (p.ej. Moodle) generan grandes cantidades de datos como resultado de las actividades realizadas por los estudiantes, de su actividad, de sus accesos y experiencia web, datos que las plataformas recogen y almacenan, pero que no se emplean en última instancia como cabría esperar para abordar el problema del abandono antes mencionado (Beck y Woolf, 2000). Estos datos son potencialmente útiles para detectar cualquier problema de seguimiento en las aulas, y en última instancia, para evitar el abandono estudiantil a través de una acción tutorial adecuada que llegue en el momento preciso al estudiante (Ben-naim *et al.,* 2008; Huang *et al.,* 2007; Matsuda *et al.,* 2007).

Cita sugerida:
Lizcano, D. (2019). Modelo de Predicción y Prevención del abandono estudiantil en enseñanzas a distancia basado en minado web. En A. Cotán Fernández (Coord.), *Nuevos paradigmas en los procesos de enseñanza-aprendizaje.* (pp. 58-65). Eindhoven, NL: Adaya Press.

Desarrollo y Metodología

En este trabajo se emplean técnicas de descubrimiento de conocimiento en bases de datos (KDD) para analizar los datos históricos de calificaciones (Baker y Yacef, 2009; Castro *et al.*, 2007; Fayyad *et al.,* 1996; Lara *et al*., 2014), acceso y uso web de cursos pasados de asignaturas de Ingeniería Informática (Redes de Computadores y Sistemas Distribuidos).

El modelo propuesto realiza una adquisición automática de información para cada asignatura (Burr y Spennemann, 2004; Ueno y Nagaoka, 2002; Vee *et al.,* 2006), basada en las distintas semanas de actividad en que se organice la misma (Beal y Cohen, 2008; Huang *et al.,* 2009; Romero y Ventura, 2010). Gracias a esa recolección, es sencillo extraer una matriz con los accesos de cada estudiante a cada recurso en cada semana, así como las calificaciones obtenidas en las distintas actividades evaluables (Chen *et al.,* 2000; Ibrahim y Rusli, 2007; Pardos *et al.,* 2007; Ritter *et al.*, 2009). Además de esta información presente en la plataforma, gracias a las actas definitivas es posible tener registro cruzado de la calificación definitiva de cada estudiante tras realizar el examen final y saber si abandono o no el curso, y en qué momento lo hizo (Chen *et al.*, 2004; Chen *et al.*, 2007; Lau *et al.*, 2007; Nebot *et al.*, 2006; Spacco *et al.,* 2006).

Esta información, registrada en forma de series temporales complejas, permite realizar un análisis, gracias al sistema predictivo desarrollado (Figura 1), y crear así un modelo de referencia estructuralmente complejo con el fin de predecir en tiempo real si un estudiante tiene peligro de abandono o no en cada momento puntual del curso en que realicemos la consulta (Baruque *et al.,* 2007; Beck y Woolf, 2000; Delgado *et al.*, 2006; Yang *et al.*, 2002).

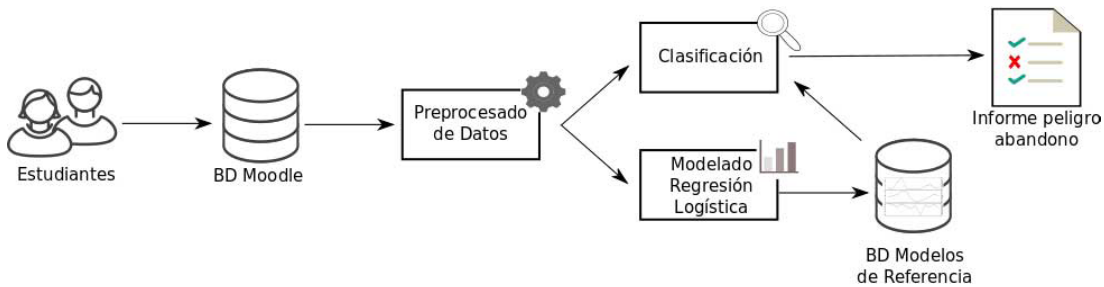

Figura 1. Visión general de la arquitectura del sistema

Para la clasificación y creación de modelos se utilizan modelos de regresión logística (Ecuación 1) (Mclaren *et al.*, 2004; Romero *et al.*, 2008).

$$P = \frac{}{1+ e^{-(w0a0,w1a1+...+wkak)}} \qquad (1)$$

En un primer lugar se realiza la creación de los modelos, siguiendo el siguiente algoritmo (1):

ALGORITMO 1

Entradas:

- *P*, tabla que contiene el horario de enseñanza especificando las presentaciones de actividades.
- *M*, tabla que contiene las calificaciones (preprocesadas) históricas de los estudiantes.
- *w*, número de semana.
- *a*, actividad de una semana.

Salidas: Coeficientes del modelo de regresión M*i*

Pasos:

1) Filtrar todas las columnas de *M* relacionadas con actividades con una fecha de vencimiento de una semana mayor que *w*

2) Construir el modelo de regresión logística M*i* basado en la Ecuación (1), donde i es la actividad calificada más recientemente en la semana *w*

3) Resolver la ecuación (1) para devolver los coeficientes utilizando el método Newton-Raphson.

4) Almacenar M*i* en la base de datos de modelos

Una vez creados los modelos con un número de datos histórico suficientemente amplio, el modelo cuenta con información suficiente como para poder clasificar nuevos estudiantes de esas asignaturas y poder predecir su posible abandono, así como el periodo temporal aproximado en el que podría producirse (Hübscher *et al.,* 2007; Ypma, 1995; Yu *et al.,* 1999). La clasificación de cada estudiante requiere la puesta en marcha del siguiente algoritmo (2):

ALGORITMO 2

Entradas:

- *w*, número de semana
- M_i, modelo de regresión de la semana *w*, donde i es la actividad más reciente de la semana w.
- *m*, lista (pre-procesada) de las actividades del estudiante a clasificar.
- a, actividad de una semana

Salidas: *Clase*: 1 = abandono, 0 = no-abandono

Pasos:

1) Seleccione de m sólo la información relativa a la calificación de las actividades cuya semana de presentación es menor o igual a w (actividades ya calificadas en la semana w).

2) Sustituir estas notas en las variables del modelo de regresión Mi (dando un valor r)

3) Si r es menor o igual a 0.5, Clase = 0

De lo contrario, Clase = 1

4) Devolver Clase

Los experimentos realizados con datos de más de 200 alumnos en varios cursos reales de formación a distancia permiten crear un modelo clasificatorio que confirma el poder predictivo de la propuesta en términos de precisión (98,3%), que supera a otros enfoques existentes.

Resultados y discusión

Utilizando los modelos predictivos obtenidos se ha diseñado un plan de acción tutorial basado en la intervención del profesorado en momentos puntuales del semestre (según la semana del curso) con el objetivo de recuperar a aquellos estudiantes que, según dichos modelos, se encontraban en riesgo de abandono escolar (Park y Choi, 2009). Aplicando dicho plan correctivo, apoyado en los modelos obtenidos, se ha conseguido reducir el abandono en un 48% frente a cursos anteriores en las asignaturas mencionadas (Figura 2).

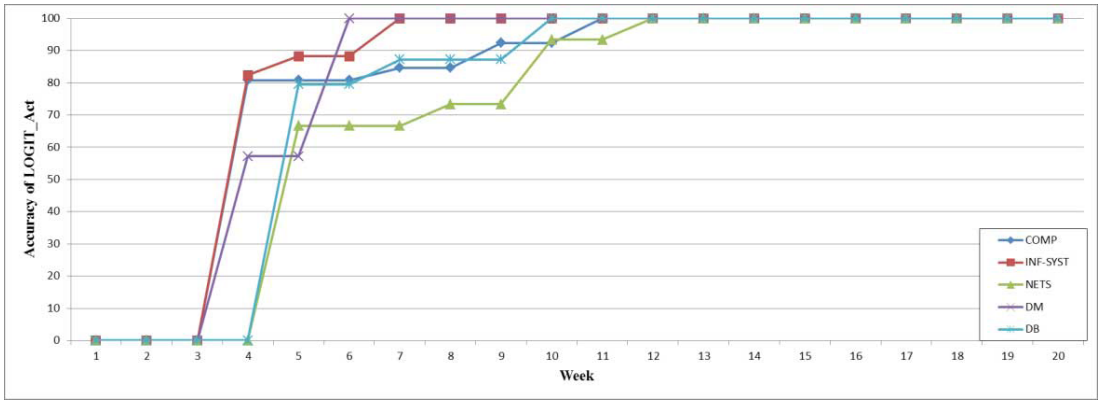

Figura. 2. Eficacia del método de prevención de abandono por semanas

Esta propuesta, por tanto, permite predecir con mucha precisión el posible abandono estudiantil en tiempo real. Así mismo es aplicable a casi cualquier entorno de aprendizaje en línea, ya que todos ellos mantienen un registro de las calificaciones de actividad de los estudiantes y son capaces de establecer y consultar las fechas de presentación. Además, la mayoría tiene una interfaz para consultar y/o exportar dichos datos (Park y Choi, 2009; Ypma, 2005). Por lo tanto, esta propuesta podría aplicarse en otros entornos y, por supuesto, a otros cursos y otras ramas de conocimiento.

Quien utilice esta propuesta debe tener en cuenta que, aunque a medio y largo plazo durante el curso los resultados han sido muy prometedores, el método no proporciona ningún resultado en las primeras semanas de actividad, es decir, el sistema alcanza una precisión superior al 90% después de que se haya presentado y corregido la primera actividad. En las primeras semanas del curso, deberían utilizarse otras técnicas de seguimiento más o menos automatizadas para complementar la propuesta presentada.

En cuanto a la interpretabilidad de los modelos, es cierto que la regresión no genera los modelos de EDM más interpretables. Sin embargo, los coeficientes del modelo asociados a cada actividad, y especialmente su valor absoluto y su signo, son analizables.

En principio, al ser variables positivas, un signo de coeficiente negativo aumentará que el denominador de la ecuación (1) y, por tanto, los resultados finales se inclinarán más hacia 0 (no abandono). Lo contrario se aplica a los coeficientes positivos. Por otro lado, el valor absoluto de cada coeficiente da una idea de la importancia de cada actividad para predecir el abandono. Por lo tanto, se pueden señalar actividades más discriminatorias basadas en su peso en los modelos de referencia.

Una vez conocidas las actividades clave de cada curso, se podría diseñar un plan de acción de tutoría a medida para cada uno de ellos, centrándose en proporcionar apoyo a las competencias aprendidas a través de estas actividades clave. Dependiendo del tipo de actividad (test, tarea práctica o trabajo colaborativo), este plan de tutoría puede incluir acciones de tutoría como sesiones extra de videoconferencia, recomendaciones para repetir la actividad, etc.

Conclusiones

En este trabajo se ha desarrollado un sistema de descubrimiento de conocimiento basado en datos académicos recopilados de una plataforma de aprendizaje en línea como Moodle. En particular, estos datos son las calificaciones históricas de los estudiantes para las diferentes actividades establecidas durante el curso analizado. El sistema diseñado es capaz de predecir, en tiempo real, si los estudiantes abandonarán o no el curso en cualquier momento, basándose en sus calificaciones y en los modelos históricos de los estudiantes anteriores. Los modelos históricos se construyeron utilizando técnicas de regresión logística.

La propuesta, basada en las calificaciones de los estudiantes, ha demostrado resultados ligeramente mejores que las propuestas existentes en términos de precisión, especialmente en las semanas cruciales del semestre (en este caso, las semanas 9 a 13 de los cursos de 20 semanas). Esto es corroborado por los experimentos realizados con más de 100 estudiantes reales de cinco cursos académicos. Los modelos obtenidos, combinados con un plan específico de acción tutorial, han resultado de gran utilidad para reducir la tasa de abandono del curso 2014-2015 en las materias estudiadas, frente a cursos previos en los que no se implementó ningún mecanismo de este tipo.

Aparte de los buenos resultados en términos de poder predictivo y eficiencia en la prevención del abandono, la propuesta es altamente aplicable independientemente de la plataforma específica de e-learning, ya que los registros de calificaciones y los horarios de enseñanza (fechas de presentación de actividades) son información básica que todos los sistemas suministran. Por otro lado, los modelos resultantes también son capaces de analizar la importancia de cada actividad y su relación con la tasa de abandono de los estudiantes en un curso en particular.

Una vez identificada la importancia relativa de cada actividad, se está considerando, como línea de investigación futura, la posibilidad de mejorar el plan de acción tutorial actual, incluyendo actividades de apoyo a medida para estas actividades clave. En la actualidad, el mecanismo propuesto se ejecuta a petición del profesorado en los momentos

que éste determina. Quizá sería interesante que esos momentos fueran determinados de forma automática. Por eso, otra línea de investigación sería el despliegue de un sistema integrado de alerta. Si se detectara una posible deserción escolar, este sistema enviaría automáticamente una copia de la advertencia al buzón de correo electrónico del tutor correspondiente. El tutor podrá entonces tomar rápida y eficazmente las medidas adecuadas descritas en el plan de acción de tutoría especificado anteriormente.

Referencias

Baker, R., y Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *J. Educ. Data Mining, 1*(1), 3–17.

Baruque, C. B., Amaral, M. A., Barcellos, A., Da Silva Freitas, J. C., y Longo, C. J. (2007). Analysing users' access logs in Moodle to improve e learning. *Proc. Euro Amer. Conf. Telematics Inf. Syst.*, Faro, Portugal, pp. 1–4.

Beal, C. R., y Cohen, P. R. (2008). Temporal data mining for educational applications. *Proc. 10th Pacific Rim Int. Conf. Artif. Intell.: Trends Artif. Intell.*, Hanoi, Vietnam, pp. 66–77.

Beck, J. E., y Woolf, B. P. (2000). High-level student modeling with machine learning. *Proc. 5th Int. Conf. Intell. Tutoring Syst.*, Alagoas, Brazil, pp. 584–593.

Ben-naim, D., Marcus, N., y Bain, M. (2008). Visualization and analysis of student interaction in an adaptive exploratory learning environment. *Proc. Int. Workshop Intell. Support Exploratory Environ. Eur. Conf. Technol. Enhanced Learn*., Maastricht, The Netherlands, pp. 1–10.

Burr, L., y Spennemann, D. H. (2004). Pattern of user behaviour in university online forums. *Int. J. Instruct. Technol. Distance Learn., 1*(10), 11–28.

Castro, F., Vellido, A., Nebot, A., y Mugica, F. (2007). Applying data mining techniques to e-learning problems. Evolution of Teaching and Learning Paradigms in Intelligent Environment. *Studies in Computational Intelligence, I.*

Chen, C., Chen, M., y Li, Y. (2007). Mining key formative assessment rules based on learner portfiles for web-based learning systems. *Proc. IEEE Int. Conf. Adv. Learn. Technol.,* Niigata, Japan, pp. 1–5.

Chen, C., Duh, L., y Liu, C. (2004). A personalized courseware recommendation system based on fuzzy item response theory. *Proc. IEEE Int. Conf. E-Technol., E-Commerce E-Service*, Washington, DC, pp. 305–308.

Chen, G., Liu, C., Ou, K., y Liu, B. (2000). Discovering decision knowledge from web log portfolio for managing classroom processes by applying decision tree and data cube technology. *J. Educ. Comput. Res., 23*(3), 305–332.

Delgado, M., Gibaja, E. Pegalajar, M. C., y Pérez, O. (2006). Predicting students' marks from. Moodle logs using neural network models. *Proc. Int. Conf. Current Dev. Technol.-Assist. Educ.*, Seville, Spain, pp. 586–590.

Fayyad, U. M., Piatetsky-Shapiro, G., y Smyth, P. (1996). From Data Mining To Knowledge Discovery: An Overview. En U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, y R. Uthurusamy (Eds.), *Advances In Knowledge Discovery And Data Mining.* (pp. 1-34). AAAI Press/The MIT Press, Menlo Park, CA.

Hershkovitz, A., y Nachmias, R. (2008). Developing a log-based motivation measuring tool. *Proc. 1st Int. Conf. Educ. Data Mining,* Montreal, QC, Canada, pp. 226–233.

Huang, C., Lin, W., Wang, W., y Wang, W. (2009). Planning of educational training courses by data mining: Using China Motor Corporation as an example. *Expert Syst. Appl. J., 36*(3), 7199–7209.

Huang, J., Zhu, A., y Luo, Q. (2007). Personality mining method in web based education system using data mining. *Proc. IEEE Int. Conf. Grey Syst. Intell. Services,* Nanjing, China, pp. 155–158.

Hübscher, R., Puntambekar, S., y Nye, A. (2007). Domain specific interactive data mining. *Proc. 11th Int. Conf. User Model. Workshop Data Mining User Model,* Corfu, Greece, pp. 81–90.

Ibrahim, Z., y Rusli, D. (2007). Predicting students' academic performance: Comparing artificial neural network, decision tree and linear regression. *Proc. Annu. SAS Malaysia Forum*, Kuala Lumpur, Malaysia, pp. 1–6.

Kotsiantis, S., Pierrakeas, C., y Pintelas, P. (2003). Preventing student dropout in distance learning systems using machine learning techniques. *Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, Oxford, U.K., pp. 3–5.

Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., y Riera, T. (2014). A system for knowledge discovery in e-learning environments within the European Higher Education Area – Application to student data from Open University of Madrid, UDIMA. *Computers & Education, 72*, 23-36.

Lau, R., Chung, A., Song, D., y Huang, Q. (2007). Towards fuzzy domain ontology based concept map generation for e-learning. *Proc. Int. Conf. Web-Based Learning,* Edinburgh, U.K., pp. 90–101.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., y Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Comput. Educ. J., 53*(3), 950–965.

Madhyastha, T., y Hunt, E. (2009). Mining diagnostic assessment data for concept similarity. *J. Educ. Data Mining, 1*(1), 72–91.

Matsuda, N., Cohen, W., Sewall, J., Lacerda, G., y Koedinger, K. R. (2007). Predicting students performance with SimStudent that learns cognitive skills from observation. *Proc. Int. Conf. Artif. Intell. Educ.,* Amsterdam, The Netherlands, pp. 467–476.

Mclaren, B. M., Koedinger, K. R., Schneider, M., Harrer, A., y Lollen, L. (2004). Bootstrapping novice data: Semi-automated tutor authoring using student log files. P*roc. Workshop Analyzing Student-Tutor Interaction Logs Improve Educ. Outcomes*, Alagoas, Brazil, pp. 1–10.

Nebot, A., Castro, F., Vellido, A., y Mugica, F. (2006). Identification of fuzzy models to predict students performance in an e-learning environment. *Proc. Int. Conf. Web-Based Educ.,* Puerto Vallarta, Mexico, pp. 74–79.

Pardos, Z., Heffernan, N., Anderson, B., y Heffernan, C. (2007). The effect of model granularity on student performance prediction using Bayesian networks. *Proc. Int. Conf. User Model,* Corfu, Greece, pp. 435–439.

Park, J.H., y Choi, H. J. (2009). Factors Influencing Adult Learners' Decision to Drop Out or Persist in Online Learning. *Educational Technology & Society, 12*(4), 207-217.

Ritter, S., Harris, T., Nixon, T., Dickison, D., Murray, R., y Towle, B. (2009). Reducing the knowledge tracing space. *Proc. Int. Conf. Educ. Data Mining,* Cordoba, Spain, pp. 151–160.

Romero, C., y Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Trans. Systems, Man and Cybernetics, Part C: Applications and Reviews, 40*(6), 601-618.

Romero, C., Ventura, S., Hervás, C., y Gonzales, P. (2008). Data mining algorithms to classify students. *Proc. Int. Conf. Educ. Data Mining,* Montreal, Canada, pp. 8–17.

Spacco, J., Winters, T., y Payne, T. (2006). Inferring use cases from unit testing. *Proc. Workshop Educ. Data Mining*, New York, pp. 1– 7.

Ueno, M., y Nagaoka, K. (2002). Learning log database and data mining system for e-learning—on line statistical outlier detection of irregular learning processes. *Proc. Int. Conf. Adv. Learning Technol.*, Tatarstan, Russia, pp. 436–438.

Vee, M. N., Meyer, B., y Mannock, K. L. (2006). Understanding novice errors and error paths in Object-oriented programming through log analysis. *Proc. Workshop Educ. Data Mining,* Taiwan, pp. 13–20.

Yang, T. D., Lin, T., y Wu, K. (2002). An agent-based recommender system for lesson plan sequencing. *Proc. Int. Conf. Adv. Learning Technol.*, Kazan, Russia, pp. 14–20.

Ypma, T. J. (1995). Historical development of the Newton-Raphson method. *SIAM Review, 37*(4), 531–551.

Yu, C. H., Jannasch-Pennell, A., Digangi, S., y Wasson, B. (1999). Using online interactive statistics for evaluating web-based instruction. *J. Educ. Media Int., 35*, 157–161.

David Lizcano Casas (D. Lizcano), es doctor en Ingeniería Informática por la Universidad Politécnica de Madrid desde 2010. Desde 2011 es profesor Titular en la Universidad a Distancia de Madrid, de la que actualmente es Vicerrector de Investigación y Doctorado. Ha realizado estancias de investigación en instituciones internacionales (como en el MIT – Media Lab), ha publicado más de 25 artículos en revistas de alto impacto (JCR), ha participado en más de 20 proyectos nacionales y europeos (FP7 y H2020) y en más de 40 congresos internacionales.

ISBN 978-94-92805-12-6