Determinación de perfiles de estudiantes y de rendimiento académico mediante la utilización de minería de datos en la UTN - FRRe

David L. la Red Martínez, Marcelo Karanik, Mirtha E. Giovannini

Grupo de Investigación Educativa / Departamento de Ingeniería en Sistemas de Información / Facultad Regional Resistencia / Universidad Tecnológica Nacional
French 414, (3500) Resistencia, Argentina, +54-379-4638194
laredmartinez@gigared.com mkaranik@gmail.com meg_c51@yahoo.com.ar

Resumen

Durante el cursado de la carrera Ingeniería en Sistemas de Información de la Facultad Regional Resistencia de la Universidad Tecnológica Nacional (UTN-FRRe) los alumnos deben ajustarse a un régimen de correlatividades, para poder cursar asignaturas de años posteriores. En este contexto, existen materias que son consideradas críticas ya que pueden provocar retrasos en el normal desempeño del alumno. Una de esas materias es Algoritmos y Estructuras de Datos que plantea desafíos relacionados con la lógica de la programación. Actualmente, la cantidad de alumnos que regularizan y aprueban la asignatura es considerablemente baja, aportando al desgranamiento y deserción en los primeros niveles. Esto denota, por tanto, la importancia de determinar cuáles son las variables que inciden en el rendimiento académico y así, establecer estrategias que permitan mejorarlo. Este trabajo presenta un modelo que incluye: a) situación del alumno: educación media, nivel educacional de los padres, educación secundaria, nivel socio-económico, edad, género, si trabaja y la actitud hacia el estudio; b) el contexto educativo: cursillo de ingreso, régimen de cursado, herramientas de apoyo académico. El modelo que se describe propone utilizar técnicas de almacenes de datos y de minería de datos, para establecer perfiles de los alumnos y determinar situaciones potenciales de éxito o de fracaso académico, para establecer acciones tendientes a evitar estos últimos.

Palabras clave: perfiles; rendimiento académico; almacenes de datos; minería de datos.

Contexto

La investigación mencionada precedentemente se encuadra en el proyecto homónimo, aprobado el 28/09/2012 por Disposición N° 37/13 de la Secretaría de Ciencia, Tecnología y Posgrado de la UTN. Dicho proyecto, financiado totalmente por la UTN, tiene el código 25/L059 en el Programa de Incentivos a los Docentes Investigadores y el código UTI1719 en el ámbito de la UTN. La vigencia del proyecto es desde el 01/01/2013 al 31/12/2015.

Introducción

Según (Joyanes Aguilar, 1997), la nueva sociedad de la información o ciber-

sociedad plantea un gran número de interrogantes de orden técnico, económico, sociológico, cultural y político.

Uno de los interrogantes es si los sistemas educativos serán capaces de producir la cantidad y calidad de egresados necesarios para soportar las demandas de personal altamente capacitado de esta sociedad de la información y el conocimiento (SIC) en las diferentes áreas, especialmente en las relacionadas con las TICs. Es acá donde aparece el problema del rendimiento o desempeño académico. En (Forteza, 1975) se define el rendimiento académico como la productividad del sujeto, matizado por sus actividades, rasgos y la percepción más o menos correcta de los cometidos asignados. No obstante, a la hora de operativizar el rendimiento, se tiende al reduccionismo (González, 1988). En (Marreno & Espino, 1988) se analiza el poder predictivo de las distintas aptitudes, mediante regresión múltiple, concluyendo que la más importante predictora del rendimiento académico es la verbal, seguida de la aptitud numérica y del razonamiento.

En un estudio acerca del rendimiento académico en el primer curso universitario (García & San Segundo, 2001) se utilizan indicadores como las tasas de graduación, diferenciando por tipos de centros y analizando el rendimiento académico a partir de datos individuales. También se ha estudiado el rendimiento académico universitario de los alumnos a través de las calificaciones de entrada a la Universidad (Vivo Molina et al., 2004), realizando el análisis de los datos mediante la técnica estadística de curva ROC (Receiver Operating Characteristic).

Asimismo, la interacción del autoconcepto y el rendimiento académico en un contexto pluricultural se estudia en (Herrera Clavero et al., 2004). Estos autores han considerado que desde las primeras investigaciones sobre el aprendizaje los estudios se centraron exclusivamente en los aspectos cognitivos; luego los investigadores descubrieron la importancia de los componentes afectivos y su influencia decisiva en el aprendizaje; finalmente se conjugaron los aspectos cognitivos y los afectivos, naciendo así el constructo llamado aprendizaje autorregulado (self-regulated learning).

También se ha estudiado el rendimiento académico universitario (Di Gresia, 2007), aplicando el enfoque de función de producción para estimar los determinantes del rendimiento académico.

En (Delfino, 1989) se han analizado los determinantes del aprendizaje mediante un enfoque de función de producción sugiriendo que los rendimientos escolares dependen de factores genéticos y socioeconómicos, de la calidad del docente, de las condiciones de la escuela y del grupo de alumnos (peer effect). Los resultados publicados en (Maradona y Calderón, 2004) han mostrado que el factor más significativamente relacionado con la calidad educativa es el propio alumno como co-productor, medido a través del nivel socioeconómico del hogar de donde proviene. En (Porto & Di Gresia, 2000) se ha mostrado que la productividad del estudiante es mayor para las mujeres, para los estudiantes de menor edad y para quienes provienen de hogares con padres más educados. En (Fazio, 2004) se ha analizado detalladamente la vinculación entre horas trabajadas y rendimiento académico.

En general los estudios empíricos confirman la correlación entre mayores niveles de educación y atributos positivos luego de los estudios (McMahon, 2002).

En California (USA), el Academic Performance Index Reports incluye aspectos relacionados al rendimiento académico (academic performance) (California Department of Education, 2010). En (Reyes, 2004) se ha mostrado el contraste

que hay entre las personas que trabajan y estudian y las que solamente estudian, encontrándose que no existen diferencias en el rendimiento académico de los dos conjuntos.

En (García Jiménez et al., 2000) se ha estudiado la capacidad de la regresión lineal y de la regresión logística en la predicción del rendimiento y del éxito/fracaso académico, partiendo de variables como la asistencia y la participación en clase. Han concluido que el rendimiento previo es un buen predictor del rendimiento futuro y que la asistencia y sobre todo la participación son variables con un peso importante en la predicción del rendimiento.

En (Marcelo, Villarín & Bermejo, 1987) se ha demostrado que las variables planificación del estudio, inteligencia, apoyo del profesor, estudio, tiempo, condiciones ambientales de estudio e implicación formaban parte de la ecuación de predicción de regresión múltiple, explicando un 25,70% de la varianza del rendimiento escolar en cursos de bachillerato.

El problema de encontrar buenos predictores del rendimiento futuro de manera que se reduzca el fracaso académico en los programas de postgrado ha recibido una especial atención en EE.UU. (Wilson & Hardgrave, 1995), habiéndose encontrado que las técnicas de clasificación como el análisis discriminante o la regresión logística son más adecuadas que la regresión lineal múltiple a la hora de predecir el éxito/fracaso académico.

Además de las metodologías tradicionales antes señalas utilizadas para el estudio del rendimiento académico, existen otras provenientes de la Inteligencia de Negocios (BI: Business Intelligence), tales como los Almacenes de Datos (Data Warehouses: DW) y la Minería de Datos (Data Mining: DM), utilizada para el descubrimiento de conocimiento oculto en grandes volúmenes de datos.

Un DW es una colección de datos orientado a temas, integrado, no volátil, de tiempo variante, que se usa para el soporte del proceso de toma de decisiones gerenciales (Inmon, 1992), (Inmon, 1996), (Simon, 1997).

La DM es la etapa de descubrimiento en el proceso de KDD (Knowledge Discovery from Databases), es el paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos pre-procesados (Fayyad, Grinstein & Wierse, 2001), (Hand, Mannila & Smyth, 2000).

La DM está muy ligada a los DW ya que los mismos proporcionan la información histórica con la cual los algoritmos de minería obtienen la información necesaria para la toma de decisiones (IBM Software Group, 2003).

La DM es un conjunto de técnica de análisis de datos que permiten extraer patrones, tendencias y regularidades para describir y comprender mejor los datos y extraer patrones y tendencias para predecir comportamientos futuros (Simon, 1997), (Berson & Smith, 1997), (White, 2001).

La DM genera modelos que pueden ser descriptivos o predictivos (Agrawal & Shafer, 1996); sus técnicas son diversas, una de las más utilizadas es la de clustering (o agrupamiento de datos) (Grabmeier, & Rudolph, 1998), (Ballard, Rollins, Ramos, Perkins, Hale, Dorneich, Cas Milner & Chodagam, 2007). El cluster demográfico es un algoritmo desarrollado por IBM, que resuelve automáticamente los problemas de definición de métricas de distancia / similitud, proporcionando criterios para definir una segmentación óptima.

Líneas de Investigación, Desarrollo e Innovación

Se intenta determinar en qué medida el desigual rendimiento académico de los alumnos de Algoritmos y Estructuras de Datos es influenciado por las siguientes variables:

a. escuela media de procedencia;
b. nivel educativo de los padres;
c. nivel socio-económico;
d. edad;
e. género;
f. actitud general hacia el estudio;
g. existencia del cursillo de ingreso;
h. régimen de cursado (anual - cuatrimestral);
i. uso de herramientas de apoyo (campus virtual).

Se busca encontrar perfiles de alumnos de rendimiento académico bajo, medio y alto utilizando minería de datos sobre un almacén de datos (La Red Martínez et al., 2010, 2012).

Actualmente se desarrolla la etapa de carga de datos para crear el almacén de datos sobre el que se aplicarán los algoritmos de DM.

Resultados y Objetivos

El objetivo general es determinar, utilizando técnicas de DW y DM, las variables que explican el desigual rendimiento académico por parte de los alumnos de Algoritmos y Estructuras de datos de la carrera de Ingeniería en Sistemas de Información de la UTN-FRRe, a fin de establecer acciones que permitan mejorarlo. En este contexto se considera rendimiento académico a los resultados logrados en las evaluaciones realizadas durante el cursado de la asignatura. Los objetivos específicos son: a) relevar información de la situación actual respecto al rendimiento académico de los alumnos de la materia enunciada en el objetivo general, b) filtrar y depurar la información contenida en las bases de datos actuales, c) establecer las variables relevantes para describir la situación objeto de estudio, d) determinar cómo influye cada una de las variables que se fijaron para evaluar la situación del alumno, e) determinar cómo influye cada una de las variables que se fijaron para evaluar el contexto académico, f) establecer acciones que tiendan a mejorar los índices de rendimiento académico de los alumnos.

Formación de Recursos Humanos

El equipo de trabajo está integrado por un Director (Doctor, Categoría II P.I., Categoría A UTN), un Co-director (Doctor, Categoría IV P.I., Categoría C UTN), un investigador (Doctor, Categoría V P.I., Categoría D UTN), un investigador (Especialista) realizando su tesis de maestría y dos becarios. En diciembre/2013 fue aprobada en la Universidad Nacional de Pilar, Paraguay, una tesis de maestría relacionada con la temática del proyecto de investigación y dirigida por el Director del mismo, quien además, en la actualidad, dirige una tesis de maestría temáticamente relacionada en la Universidad Nacional del Este, Paraguay y una tesina de grado en la Universidad Nacional del Nordeste, Argentina.

Referencias

Agrawal, R.; Shafer, J. C. Parallel Mining of Association Rules. IEEE Transactions on Knowledge and Data Engineering. December 1996. USA. 1996.

Ballard, Ch.; Rollins, J.; Ramos, J.; Perkins, A.; Hale, R.; Dorneich, A.; Cas Milner, E. & Chodagam, J. Dynamic Warehousing: Data Mining Made Easy. IBM International Technical Support Organization. IBM Press. USA. 2007.

Berson, A. & Smith, S. J. Data Warehouse, Data Mining & OLAP. Mc Graw Hill. USA. 1997.

California Department of Education. 2009–10 Academic Performance Index Reports. USA. 2010.

Delfino, J. A. Los determinantes del aprendizaje. In Petrei, A. H., editor, Ensayos en economía de la educación. Educational Evaluation and Policy Analisys. 1989.

Di Gresia, L. Rendimiento Académico Universitario. Tesis Doctoral. Universidad Nacional de La Plata. Argentina. 2007.

Fayyad, U.M.; Grinstein, G. & Wierse, A. Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann. Harcourt Intl. 2001.

Fazio, M. V. Incidencia de las horas trabajadas en el rendimiento académico de estudiantes universitarios argentinos. Documentos de Trabajo UNLP, 52. Argentina. 2004.

Forteza, J. Modelo instrumental de las relaciones entre variables motivacionales y rendimiento. Revista de Psicología General y Aplicada, 132, 75-91. España. 1975.

García, M. M.; San Segundo, M. J. El Rendimiento Académico en el Primer Curso Universitario. X Jornadas de la Asociación de Economía de la Educación. Libro de Actas, págs. 435-445. España. 2001.

García Jiménez, M. V.; Alvarado Izquierdo, J. M.; Jiménez Blanco, A. La predicción del rendimiento académico: regresión lineal versus regresión logística. Psicothema Vol. 12, Supl. nº 2, pp. 248-252. España. 2000.

González, A. J. Indicadores del rendimiento escolar: relación entre pruebas objetivas y calificaciones. Revista de Educación, 287, 31-54. España. 1988.

Grabmeier, J. & Rudolph, A. Techniques of Cluster Algorithms in Data Mining version 2.0. IBM Deutschland Informationssysteme GmbH. GBIS (Global Business Intelligence Solutions). Germany. 1998.

Hand, D.J.; Mannila, H. & Smyth, P. Principles of Data Mining. The MIT Press. USA. 2000.

Herrera Clavero, F. et al. ¿Cómo Interactúan el Autoconcepto y el Rendimiento Académico en un Contexto Educativo Pluricultural?. Revista Iberoamericana de Educación. España. 2004.

IBM Software Group. Enterprise Data Warehousing whit DB2: The 10 Terabyte TPC-H Benchmark. IBM Press. USA. 2003.

Inmon, W. H. Data Warehouse Performance. John Wiley & Sons. USA. 1992.

Inmon, W. H. Building the Data Warehouse. John Wiley & Sons. USA. 1996.

Joyanes Aguilar, L. Cibersociedad. Mc Graw Hill. España. 1997.

La Red Martínez, D. L.; Acosta, J. C.; Uribe, V. E.; Rambo, A. R. Academic Performance: An Approach From Data Mining. Journal of Systemics, Cybernetics and Informatics; V. 10 N° 1 2012, págs. 66-72; USA. 2012.

La Red Martínez, D. L.; Acosta, J. C.; Uribe, V. E.; Rambo, A. R.; Cutro, A. L. Data Warehouse y Data Mining Aplicados al Estudio del Rendimiento Académico. CISCI 2010 (9na. Conferencia Iberoamericana en Sistemas, Cibernética e Informática); Memorias, Volumen I, págs. 289-294; ISBN N° 978-1-934272-94-7; Orlando, Florida, USA. 2010.

Maradona, G. & Calderón, M. I. Una aplicación del enfoque de la función de producción en educación. Revista de Economía y Estadística, Universidad Nacional de Córdoba, XLII. Argentina. 2004.

Marcelo García, C.; Villarín Martínez, M.; Bermejo Campos, B. Contextualización del rendimiento en bachillerato. Revista de Educación, 282, 267-283. España. 1987.

Marreno Hernández, H.; Orlando Espino, M. Evaluación comparativa del poder predictor de las aptitudes sobre notas escolares y pruebas objetivas. Revista de Educación, 287, 97-112. España. 1988.

McMahon, W. W. Education and Development. Oxford University Press. 2002.

Porto, A. & Di Gresia, L. Características y rendimiento de estudiantes universitarios. El caso de la Facultad de Ciencias Económicas de la Universidad Nacional de La Plata. Documentos de Trabajo UNLP, 24. 2000.

Reyes R, S. L. El Bajo Rendimiento Académico de los Estudiantes Universitarios. Una Aproximación a sus Causas. Revista Theorethikos. Año VI, N° 18, Enero-Junio, 2004. El Salvador. 2004.

Simon, A. Data Warehouse, Data Mining and OLAP. John Wiley & Sons. USA. 1997.

Vivo Molina, J. M.; Franco Nicolás, M.; Sánchez de la Vega, M. del M. Estudio del rendimiento académico universitario basado en curvas ROC. Revista de Investigación Educativa, RIE, Vol. 22, Nº 2, 2004, págs. 327-340. España. 2004.

White, C. J. IBM Enterprise Analytics for the Intelligent e-Business. IBM Press. USA. 2001.

Wilson, R. L.; Hardgrave, B. C. Predicting graduate student success in an MBA program: Regression versus classification. Educational and Psychological Measurement, 55, 186-195. USA. 1995.