# Determinación de perfiles de rendimiento académico en la UNNE con Minería de Datos Educacional

Julio C. Acosta[1-2], David La Red Martínez[1] Carlos Primorac[1]

[1]Facultad de Ciencias Exactas y Naturales y Agrimensura
Universidad Nacional del Nordeste
9 de julio Nº 1449, (3400) Corrientes, Argentina.
julioaforever@hotmail.com, lrmdavid@exa.unne.edu.ar, carlosprimorac@gmail.com

[2]Facultad de Ciencias Agrarias / Universidad Nacional del Nordeste
J. B. Cabral Nº 2131, (3400) Corrientes, Argentina.

## Resumen

En este trabajo se propone evaluar el rendimiento de los estudiantes mediante técnicas de Minería de Datos. La propuesta no se enfoca en analizar el perfil del estudiante solo a través de sus calificaciones, sino también, estudiar el desempeño académico en base a otras variables.

Para definir los perfiles de los estudiantes y determinar patrones que conduzcan al éxito o fracaso académico, implementaremos un modelo que relaciona las calificaciones de los estudiantes con otras variables, tales como factores socioeconómicos, demográficos, actitudinales, entre otros; en base a lo cual clasificaremos los diferentes perfiles de alumnos.

Describimos el modelo a implementar con el uso de Data Warehouse para determinar los perfiles de rendimiento académico en las asignaturas Algebra de la carrera Licenciatura en Sistemas de Información (LSI) de la Facultad de Ciencias Exactas y Naturales y Agrimensura (FaCENA) de la Universidad Nacional del Nordeste (UNNE) y Matemática I de la carrera Ingeniería Agronómica (IA) de la Facultad de Ciencias Agrarias (FCA) de la UNNE (PI 16F002 acreditado por Res. N° 970/16 CS).

Esperamos contribuir a encontrar una respuesta al bajo rendimiento académico de los alumnos observado históricamente, problema éste que es el disparador de nuestra investigación. Los modelos predictivos que buscamos, permitirán tomar acciones tendientes a evitar el fracaso académico, detectando los alumnos con perfil de riesgo de fracaso académico de manera temprana, a poco del inicio del cursado de las asignaturas; lo que permitirá concentrar en ellos los esfuerzos de tutorías y apoyos especiales.

**Palabras clave:** rendimiento académico; almacenes de datos; minería de datos; modelos predictivos.

## Introducción

En el mejoramiento de la calidad académica en la Universidad, no necesariamente debe enfocarse sólo en el sistema de enseñanza-aprendizaje, sino que debe atender otras variables, como por ejemplo, la sistematización de procesos de evaluación permanentes que permitan monitorear cuestiones ligadas a la calidad académica y retroalimente la propuesta de mejora para la Universidad (Briand et al., 1999). Uno de los factores más críticos que debe evaluarse continuamente es el rendimiento académico. Se define al rendimiento académico como la productividad del sujeto, matizado por sus actividades, rasgos y la percepción más o menos correcta de los cometidos asignados (Maletic et al., 2002). Al evaluar el rendimiento académico se analizarán elementos que influyen en el desempeño como: los factores socioeconómicos, la amplitud de programas de estudio, las metodologías de enseñanza, los conocimientos previos del alumno (Marcus, 2003); por esto, no resulta adecuado evaluar el desempeño general de los alumnos a través de porcentajes de aprobación, notas obtenidas, etc., ya que este proceso de evaluación no brinda toda la información necesaria que pueda ser utilizada para detectar, y corregir problemas cognitivos, de aprehensión, de discernimiento, actitudinales.

Implementamos un mecanismo que nos permite determinar las características propias del estudiante analizando la existencia de patrones de comportamiento y de condiciones de los estudiantes que posibiliten la definición de los perfiles de alumnos. Actualmente existen varios métodos para determinar y clasificar patrones que se utilizan en el área de la Inteligencia Artificial y del Aprendizaje de Máquinas (del inglés Machine Learning – ML) (Marcus & Maletic, 2003). La Minería de Datos (del inglés Data Mining - DM), son procesos de descubrimiento de nuevas y significativas relaciones, patrones y tendencias en grandes volúmenes de datos utilizando técnicas de AI y

ML. Estas técnicas permiten extraer patrones y tendencias para describir y comprender mejor los datos y predecir comportamientos futuros. Un DW es una colección de datos orientada a un dominio, integrada, no volátil y variante en el tiempo para ayudar a tomar decisiones (Salton, 1989). Los DW surgieron por dos razones: a) la necesidad de proporcionar una fuente única de datos limpia y consistente para propósitos de apoyo para la toma de decisiones; b) la necesidad de hacerlo sin afectar a los sistemas operacionales (Molina López & García Herrero, 2006). En este trabajo se propone la utilización de técnicas de DM, con volúmenes no muy grandes de datos que oscilaran de cientos a miles, sobre información del desempeño de los alumnos de las cátedras Algebra (LSI) FaCENA-UNNE y Matemática I de la FCA-UNNE.

## Materiales y métodos

Trabajamos para detectar grupos de estudiantes en riesgo de fracaso en sus estudios, a fin de adoptar acciones proactivas frente al desgranamiento y el bajo rendimiento académico de los alumnos de primer año en la Universidad. La experiencia se realiza en las asignaturas Algebra de la carrera LSI de la FaCENA de la UNNE y en Matemática I de la carrera IA de la FCA de la UNNE.

Si bien ambas asignaturas tienen régimen de acreditación similar, difieren en la carga horaria y los tiempos de dictado a saber: Algebra (LSI) tiene 128 (ciento veintiocho) horas reloj de dictado de las cuales el 50% corresponde a teoría y el 50 % a trabajos prácticos en la modalidad cuatrimestral (corresponde al primer cuatrimestre de primer año de la carrera), mientras Matemática I (IA) tiene 96 (noventa y seis) horas reloj de dictado con idéntica distribución porcentual de timpos de dictado de teoría y de trabajos prácticos, pero en la modalidad trimestral (corresponde al primer trimestre de primer año de la carrera)

En ambas asignaturas para alcanzar la condición de alumno regular, los alumnos deben asistir al menos al 75% de las clases de trabajos prácticos, que se dictan dos veces por semana en clases de 2 hs. cada una y deben aprobar 2 (dos) exámenes parciales cuyos contenidos son exclusivamente de trabajos prácticos; cada uno de ellos tiene su instancia de recuperación y para aquellos alumnos que hayan aprobado al menos 1 (uno) de los parciales en cualquiera de las 4 (cuatro) instancias disponibles, existe una instancia más para recuperar el examen que queda aún sin aprobar. Cualquiera de los exámenes parciales se aprueba con 60 (sesenta) puntos sobre 100 (cien) puntos posibles. La asistencia a clases de teoría es libre y se dictan dos veces por semana en clases de 2 hs. cada una.

Se acreditan las asignaturas con un examen final al que se accede en condición de alumno regular o de alumno libre; el alumno regular debe rendir en el examen final solamente los contenidos de teoría en un examen oral. El alumno que se presenta al examen final en condición de alumno libre, debe rendir un examen escrito de trabajos prácticos y tras aprobar esa instancia pasa al examen de teoría en condiciones similares a la antes mencionada.

Los porcentuales de los alumnos que regularizan las Algebra y Matemática I no son los deseados; en el caso de Algebra, de 320 alumnos inscriptos en los últimos 4 años, aproximadamente un 20% no alcanza a rendir el primer examen parcial en promedio y al final del cursado, regularizan la asignatura solo un 30% aproximadamente, en el caso de Matemática I el desgranamiento después del primer parcial no es tan evidente y el porcentual aproximado de alumnos regulares al final del cursado es del 40%.

La cantidad de alumnos que regularizan y/o que aprueban las asignaturas involucradas en este proyecto no es satisfactoria, consideramos que esa situación puede contribuir al desgranamiento y deserción de los alumnos en los primeros niveles de sus carreras. Es importante, por tanto, estudiar y determinar cuáles son las variables que inciden en el rendimiento académico a fin de poder establecer estrategias de acción pedagógicas que permitan mejorar dicho rendimiento.

Trabajaremos principalmente en el desarrollo de métodos que contribuyan a encontrar técnicas para la detección temprana de los alumnos que tendrán dificultades en sus estudios, a fin ofrecerles una contención y acompañamiento especial en el inicio de sus estudios Universitarios. Indagaremos aspectos tales como: a) diferencia del nivel de aprendizajes de contenidos previos en los alumnos, b) situaciones particulares personales de los propios alumnos, c) la capacidad de las cátedras para el seguimiento del aprendizaje de los alumnos, d) escasa motivación para el estudio de ciencias básicas y otros que puedan revelarse como incidentes en la problemática que nos ocupa y otros que serán detallados adelante.

Para recuperar contenidos en los grupos de riesgo detectados trabajaremos con materiales elaborados con nuevas tecnologías de la informacion (NTIC). Esto no debe desplazar ni sustituir las formas presenciales de enseñanza - aprendizaje, sino más bien situarnos en la posición de ofrecer alternativas diferentes para aquellos alumnos que requieren modelos diferentes para sus estudios y aprendizajes. Consideramos que las NTIC tienen el potencial para desempeñar un papel importante en la recuperación de contenidos al permitir un

abordaje más eficaz, en el sentido de permitirnos procesos de aprendizaje más profundos y más persistentes (Motschnig-Pitrik & Holzinger, 2002), mientras el peso de un aprendizaje efectivo permanece con las personas, sus capacidades y valores interpersonales (Derntl et al., 2011). En tal sentido, entendemos importante en nuestro trabajo el estudio que se efectuará en dos poblaciones aparentemente diferentes como son los alumnos de las carreras Licenciatura en Sistemas de Información y los alumnos de Ingeniería Agronómica, para determinar si los perfiles de los estudiantes varían según la elección de la carrera y medir las diferencias en la predisposición y adaptación para el trabajo y aprendizaje mediado con las NTICs (lo cual se confirmará o no).

En los últimos años se han realizado numerosos trabajos relacionados con la producción de contenidos; actualmente se tiene una concepción global e integral del e-learning (Nichols, 2003), en estos nuevos escenarios se incluyen la combinación del aprendizaje cara a cara y el soportado por medios tecnológicos (especialmente la Web), tal que las fortalezas de ambas configuraciones se puedan aprovechar y explotar. Este aprendizaje combinado (blended learning o b-learning) se considera de suma utilidad no sólo para las universidades sino también para la sociedad en general.

Nosotros, desde nuestros trabajos previos, hemos podido corroborar lo que oportunamente hemos formulado, que los docentes del siglo XXI deben incorporar definitivamente las NTICs como recursos didácticos, sin abandonar los tradicionales de tiza y pizarrón, pero deben conocer el uso de las NTICs con al menos en parte del potencial que ellas ofrecen (Acosta & La Red Martínez, 2012); algunas teorías psicológicas y pedagógicas consideran necesaria la inclusión del e-moderator o e-moderador, docente con habilidades especiales en las actividades online (Salmon, 2000); la actividad del docente tutor se transforma a veces en un hecho fundamental, ya que la manera en que se usa la tecnología puede transformarse en un factor de gran influencia en la calidad de la EA-EV (enseñanza - aprendizaje en entornos virtuales). Se debe trabajar entonces para lograr una forma de EA-EV que tome en cuenta las necesidades individuales, los intereses y estilos (Wenger et al., 2009).

En este proyecto de investigación, las variables que inciden en el rendimiento académico de los alumnos serán detectadas a fin de establecer, a través de los valores que ellas toman en cada caso, la población de alumnos en riesgo de fracaso, para establecer acciones tendientes a evitar el fracaso de cada uno de los alumnos, con las acciones que correspondan en cada caso particular y/o de cada grupo detectado y disminuir así el posterior desgranamiento.

## Data warehouse

Como soporte de los datos trabajaremos con Data Warehouse (DW); en informática, un almacén de datos (DW), es un sistema especial de bases de datos utilizado para el almacenamiento de datos y el procesamiento de los mismos para la presentación de informes y análisis de información, es considerado como un componente central de la inteligencia de organizaciones.

Un DW es un repositorio de datos que proporciona una visión global, común e integrada de los datos (Curto Días, 2010) (Figura 1) y presenta las siguientes características: a) Orientado a un tema: organiza una colección de información alrededor de un tema central. b) Integrado: incluye datos de múltiples orígenes y presenta consistencia de datos. c) Variable con el tiempo: se realizan fotos de los datos basadas en fechas o hechos. d) No volátil: sólo de lectura para los usuarios finales.

Detrás de la arquitectura de componentes del DW existe un conjunto de procesos básicos asociados: los ETL (del inglés Extract, Transform, Load – Extracción, Transformación y Carga). Los procesos ETL hacen referencia a la recuperación y transformación de los datos desde las fuentes orígenes cargándolos en el DW. En primer lugar los datos se analizan desde las fuentes y se extraen aquellos que serán de utilidad para el proceso en ejecución.

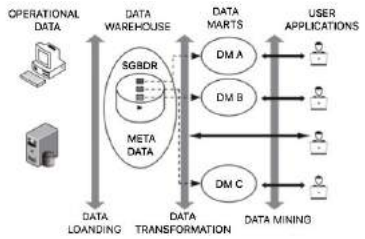


Figura 1 - Arquitectura Básica de un DW.

Luego de extraer los datos se los carga al DW pero, en muchas ocasiones, éstos requieren pasar por un proceso de transformación. La transformación de los datos significa un formateo y/o estandarización de los mismos convirtiendo ciertos números en fechas, eliminando campos nulos, etc.

Es necesario que antes de completar el DW con los datos se realicen controles para enviar información cualitativamente correcta. Luego se procede a aplicar alguna técnica para realizar el análisis de los datos almacenados en el DW. El método más utilizado es el proceso de DM que aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos

permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad.
Existen varias alternativas del DM, por ejemplo la Minería de Datos en Educación (Educational Data Mining, EDM). El objetivo de la EDM es el desarrollo de métodos para la exploración de tipos de datos únicos provenientes de plataformas educativas y usándolos para entender mejor a los estudiantes en el aprendizaje (Baker & Yaceff, 2009). Existen diversos estudios y publicaciones que abordan la evaluación de rendimiento académico utilizando técnicas de Minería de Datos (Formia & Lanzarini, 2013); (Pereira et al., 2013); (La Red Martínez et al. 2012); (La Red Martínez et al., 2017).

*Modelo propuesto*: La estructura del DW, se muestra en la Figura 2, consta de una tabla de hechos y varias tablas de dimensión. Una tabla de hechos o una entidad de hecho es una tabla o entidad que almacena medidas para medir el negocio como las ventas, el coste de las mercancías o las ganancias (IBM Knowledge Center, 2015).

Figura 2 - Modelo Propuesto del DW

Cada medida se corresponde con una intersección de valores de las dimensiones y generalmente se trata de cantidades numéricas, continuamente evaluadas y aditivas. Se pueden distinguir dos tipos de columnas en una tabla de hechos, columnas de hechos y columnas llaves. Las columnas de hechos almacenan las medidas del negocio que se quieren controlar y las columnas llaves forman parte de la clave de la tabla. Una tabla de dimensiones o entidad de dimensiones es una tabla o entidad que almacena detalles acerca de hechos. Por ejemplo una tabla de dimensión de hora almacena los distintos aspectos del tiempo como el año, trimestre, mes y día. Además incluye información descriptiva sobre los valores numéricos de una tabla de hechos. Las tablas de dimensiones para una aplicación de análisis de mercado, por ejemplo, pueden incluir el tipo de período de tiempo, región comercial y producto. Asimismo las tablas de dimensiones describen los distintos aspectos de un proceso de negocio. Si se desea determinar los objetivos de ventas, se pueden almacenar los atributos de dichos objetivos en una tabla de dimensiones. Cada tabla de dimensiones contiene una clave simple y un conjunto de atributos que describen la dimensión.
En nuestro caso, las columnas de una tabla de dimensiones se utilizan para crear informes o para mostrar resultados de consultas. Por ejemplo las descripciones textuales de un informe se crean desde las etiquetas de las columnas de una tabla de dimensiones. El modelo que se presenta en este trabajo se compone de la tabla de hechos "ALUMNOS" y varias tablas de dimensiones asociadas a la misma que incluyen características que se desean estudiar. En la Figura 2 se representa gráficamente esta estructura.

*Etapa de recolección de datos*: Tal como se planteó, el estudio del desempeño académico de los estudiantes no sólo debe evaluarse teniendo en cuenta los resultados de las instancias de evaluaciones previstas por la asignatura sino que también deben analizarse otros factores culturales, sociales y/o económicos que afecten el rendimiento del alumno. Por ello para este trabajo resultó determinante la participación directa del estudiante, pues era necesario conocer datos sobre aspectos personales que no se podían obtener de otra manera que no fuera a través de respuestas directas por parte de cada alumno. A tal fin se dispuso la elaboración de una aplicación web que permitió contar con una Encuesta On-Line compuesta por preguntas relacionadas a situación familiar e historial de estudios secundarios, entre otras cuestiones.

*Etapa de depuración y preparación de datos:* Para la realización de una correcta explotación del DW se debe asegurar que los datos obtenidos en la etapa anterior sean consistentes y mantengan la coherencia entre ellos. Así, en la etapa siguiente, se realizará un proceso de limpieza en los datos, que es la eliminación de aquellos registros con todos sus campos en blanco, corrección de errores tipográficos, llenado de algunos campos nulos, entre otros. La Encuesta no permite la carga, por parte de los estudiantes, de calificaciones de la asignatura en estudio. Esto se dispuso así para evitar errores en los datos ya sea por olvido, o confusión al momento de ingresar los valores. Por ello la carga de notas correspondientes al primer parcial, segundo parcial y sus recuperatorios, examen final y situación del alumno (regular, promovido o libre), es realizada por el equipo responsable de este trabajo de investigación. La información se obtendrá a partir de la base de datos histórica de las cátedras continuarárespecto a calificaciones de los alumnos. Con esta

información depurada se deberá proceder atrabajar en las próximas etapas: - Carga de Datos al DW: Mediante la ejecución del flujo de datos, la información almacenada en la tabla *encuesta* se distribuirá a las tablas pertenecientes al modelo del DW.

## Resultados

Hasta el momento se ha completado la primera etapa que implicó el diseño del modelo del DW sobre el cual se implementarán técnicas de DM a fin de determinar perfiles de estudiantes vinculados a su desempeño académico en las asignaturas LSI-FaCENA e IA-FCA UNNE. En el avance que aquí se presenta respecto del Proyecto se pudo comprobar que la etapa de depuración y preparación de los datos ha demandado tiempo y esfuerzo debido principalmente a la poca integridad y coherencia que existía en la información que se utilizará para realizar la evaluación final. En etapas sucesivas se continuará con el proceso de minería de datos para evaluar y comparar patrones que se obtengan para definir los perfiles de estudiantes. La evaluación, análisis y utilidad de estos patrones con los que se construirá un modelo predictivo de rendimiento académico permitirá soportar la toma de decisiones eficaces por parte del cuerpo docente de las asignaturas involucradas

## Referencias

Acosta, J. & La Red Martínez, D. (2012). Un aula virtual no convencional de Algebra en la FaCENA-UNNE. Saarbrücken: EAE.

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

Briand, L., Daly,J. & Wüst, J. (1999). A unified framework for coupling measurement in objectoriented systems. *IEEE Transactions on Software Engineering*, 25 (1), 91-121.

Curto Dias (2010). Introducción al business intelligence. UOC: Barcelona

Derntl, M., Hampel, T., Motschnig-Pitrik, R., & Pitner. (2011). Inclusive social tagging and its support in Web 2.0 services. Computers in Hum an Behavior, 27(4), 1460-1466.

Formia, S. & Lanzarini, L. (2013). Caracterización de la deserción universitaria en la UNRN utilizando minería de datos. *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología* (TE&ET), (11):92–98.

La Red Martínez, D.; Acosta J., Uribe V. & Rambo A.; (2012) Academic Performance: An Approach From Data Mining; V. 10 N° 1; Journal of Systemics, Cybernetics and Informatics; pp. 66-72; U.S.A.

La Red Martínez, D.; Giovannini, M.; Báez Molinas, M.; Torre, J. & Yaccuzzi, N. (2017) Academic performance problems: A predictive data mining-based model. Academia Journal of Educational Research; 5, 4 pp. 61-75. England, U.K.

Maletic, J., Collard, M. & Marcus, A (2002). Source Code Files as Structured Documents. *Proceedings 10th IEEE International Workshop on Program Comprehension (IWPC'02)*, 289-292.

Marcus, A. (2003). Semantic Driven Program Analysis, Kent, OH, USA, Kent State University Doctoral Thesis.

Marcus, A. & Maletic, J. (2003). Recovering Documentation-to-Source-Code Traceability Links using Latent Semantic Indexing. *Proceedings 25th IEEE/ACM International Conference on Software Engineering (ICSE'03)*. 3(10), 125-137.

Molina López, J. & García Herrero, J. (2006). *Técnicas de Análisis de Datos*. Madrid: Universidad Carlos III.

Motsching-Pitrik, R., & Holzinger, A. (2002). Student-centered teaching meets new media: concept and case study. *Journal o f E ducational Technoplogy a n d Society*, 5(4), 160-172.

Nichols, M. (2003). A theory for e-Learning. Journal o f E ducational Technology and Society, -(2), 1-10.

Salmon, G. (2000). E-moderating: The key to teaching and learning online. London: Kogan Page.

Salton, G (1989). *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Boston: Addison-Wesley Longhman Publishing Co.

Pereira, R., Romero, A. & Toledo J. (2013) Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos. *Vínculos.* (10) 1, 374-383 . Universidad distrital Francisco José de Caldas. Colombia.

Wenger, E., White, D. & Smith, J. D. (2009). Digital habitats. Stewarding technology for communities: Portland, OR, USA: Cpsquare.