Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil

Osvaldo M. Spositto
Departamento de Ingeniería e Investigaciones Tecnológicas - Universidad Nacional de La Matanza
San Justo – Provincia de Buenos Aires – Argentina
spositto@unlam.edu.ar

Martín E. Etcheverry
Departamento de Ingeniería e Investigaciones Tecnológicas - Universidad Nacional de La Matanza
San Justo – Provincia de Buenos Aires – Argentina
metcheverry@unlam.edu.ar

Hugo L. Ryckeboer
Departamento de Ingeniería e Investigaciones Tecnológicas - Universidad Nacional de La Matanza
San Justo – Provincia de Buenos Aires – Argentina
hugor@unlam.edu.ar

Julio Bossero
Departamento de Ingeniería e Investigaciones Tecnológicas - Universidad Nacional de La Matanza
San Justo – Provincia de Buenos Aires – Argentina
jbossero@unlam.edu.ar

RESUMEN

Este artículo presenta los resultados de la evaluación del rendimiento académico y de la deserción estudiantil de los estudiantes del Departamento de Ingeniería e Investigaciones Tecnológicas (DIIT) de la Universidad Nacional de La Matanza (UNLaM). La investigación se realizó aplicando el proceso de descubrimiento de conocimiento sobre los datos de alumnos del período 2003-2008. La implementación de este proceso se realizó con el software MS SQL Server para la generación de un almacén de datos, el software SPSS para realizar un preprocesamiento de los datos y el software Weka (Waikato Environment for Knowledge Analysis) para encontrar un clasificador del rendimiento académico y para detectar los patrones determinantes de la deserción estudiantil.

Palabras claves: Minería de Datos, Rendimiento Académico, Deserción Estudiantil, Almacén de Datos, Detección de Patrones.

1 INTRODUCCIÓN

Debido a la gran cantidad de información generada por las distintas áreas de cualquier institución resulta imprescindible la utilización de las Tecnologías de la Información y la Comunicación (TIC) para que la información pueda ser almacenada, transformada, analizada y visualizada. Algunas de las TIC utilizadas en este proyecto como parte del proceso de descubrimiento del conocimiento en bases de datos (DCBD) fueron: base de datos, análisis estadístico unidimensional y multidimensional y aprendizaje automático.

La UNLaM está ubicada en el Partido de La Matanza cuya cabecera es la ciudad de San Justo, tiene más de 2.100.000 habitantes proyectando al año 2009 los datos del último censo y es el distrito con más alta densidad poblacional del interior del país. La población del partido está integrada en su mayoría por clase obrera y clase media baja y en mucho menor medida clase media-media en áreas residenciales. El 79% de los estudiantes de la UNLaM habitan en el Partido y el 21% restante en las zonas de influencia (Partidos de Morón y Tres de Febrero y los barrios del Oeste de la Ciudad Autónoma de Buenos Aires).

El DIIT desarrolla desde el año 2003 un plan sistemático que incluye diferentes proyectos que, en forma articulada, intentan disminuir los índices de deserción y cronicidad ya que este es un problema lo suficientemente complejo como para abordarlo con una sola estrategia. Esta investigación forma parte del conjunto de acciones que fueron planificadas en el marco de las acreditaciones de las carreras de Ingeniería Electrónica e Ingeniería Industrial.

Las carreras que se dictan en la UNLaM están distribuidas en 4 Departamentos (Unidades Académicas) y tomando los datos del año 2008 se encuentran matriculados aproximadamente 35000 estudiantes. En el DIIT se dictan las carreras de Ingeniería Informática, Ingeniería Electrónica e Ingeniería Industrial cuyas matrículas son 4480, 919 y 613 respectivamente.

En la Tabla 1 se pueden observar los resultados de un primer análisis cuantitativo del rendimiento de los estudiantes del DIIT durante el año 2008.

Tabla 1. Rendimiento de los estudiantes desagregado por carrera.

Asignaturas Aprobadas	Cantidad de alumnos Informática	Cantidad de alumnos Electrónica	Cantidad de alumnos Industrial
0	467	199	70
1	784	106	79
2	1182	186	147
3	799	130	119
4	542	160	64
5	392	56	56
Más de 5	314	82	78
Total	4480	919	613

El objetivo de este trabajo es presentar un estudio que utilizando el proceso DCDB permita, a través de clasificadores, identificar:

- el rendimiento académico de los alumnos.
- los patrones determinantes de la deserción estudiantil.

Durante las distintas etapas de este proceso se utilizaron los datos de los alumnos desde el año 2003 hasta el año 2008. Las herramientas de software utilizadas fueron:

- el motor de base de datos MS SQL Server para realizar la recopilación, integración y almacenamiento de los datos.
- el programa estadístico SPPS para realizar la depuración, selección y transformación de los datos.
- el programa Weka para obtener los clasificadores aplicando técnicas de minería de datos.

2 TECNOLOGÍAS Y HERRAMIENTAS UTILIZADAS

2.1 Proceso de descubrimiento del conocimiento en base de datos.

El DCDB es un proceso complejo ya que no solo incluye la obtención de los modelos o patrones, sino también la evaluación e interpretación de los mismos [8]. El DCDB es definido en [4] como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, compresibles a partir de los datos". Las principales tareas del proceso de DCDB se pueden resumir en: preprocesar los datos, hacer minería de datos, evaluar los resultados y presentarlos [2][6][7][10]. En la Figura 1 se puede observar que el proceso de DCDB está organizado en 5 fases [8]:

- **recopilación e integración:** en esta fase se seleccionan las distintas fuentes de información y se transforman los datos a un formato y unidad de medida comunes generando un almacén de datos [1].
- **limpieza, selección y transformación;** en esta fase se eliminan o se corrigen los valores faltantes/erróneos y se seleccionan los atributos más relevantes o se generan nuevos atributos a partir de los existentes para reducir la complejidad de la fase de minería de datos. También se puede reducir la cantidad de instancias.
- **minería de datos:** esta es la fase donde se eligen el trabajo a realizar (clasificación, agrupamiento, etc.) y el método a utilizar.
- **evaluación e interpretación:** en este punto se analizan y evalúan los patrones obtenidos y en caso de ser necesario se retorna a alguna de las fases anteriores.
- **difusión y uso (presentación):** en esta fase se hace uso de los resultados obtenidos y se difunden entre todos los potenciales usuarios.

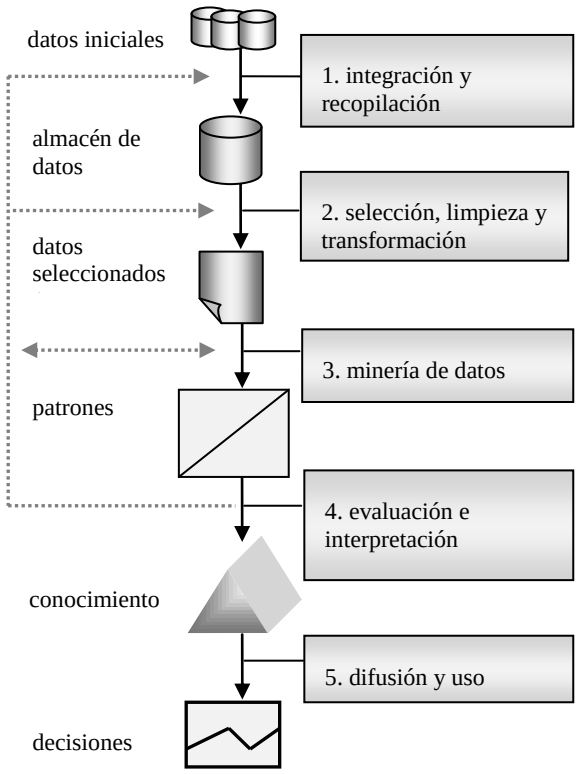

Figura 1. Fases del proceso de descubrimiento de conocimiento en bases de datos, DCDB.

2.2 Herramientas de software.

Las herramientas de software utilizadas en esta investigación fueron:

- **MS SQL Server:** se utilizó para recopilar los datos de las fuentes de información seleccionadas, para realizar una transformación de los datos a partir de la definición de los formatos y las medidas comunes y por último para almacenar los datos transformados (Almacén de Datos).
- **SPSS:** se utilizó para realizar un análisis exploratorio y de correspondencias de los datos. Como resultado del análisis se seleccionaron los atributos más relevantes y se generaron nuevos atributos a partir de los existentes.
- **Weka:** se utilizó para encontrar los patrones que permitan evaluar el rendimiento académico y la

deserción estudiantil. Este software contiene múltiples algoritmos para la aplicación de técnicas supervisadas y no supervisadas [3][9].

El software MS SQL Server es propiedad de Microsoft y el software SPPS es propiedad de IBM. El software Weka es un desarrollo de la Universidad de Waikato y se puede obtener en forma gratuita en el sitio oficial de esta institución en Internet [9].

3 RESULTADOS DEL PROCESO DE DESCUBRIMIENTO DEL CONOCIMIENTO EN BASES DE DATOS

3.1 Fase de recopilación e integración.

El resultado de esta fase fue la generación de un almacén de datos conformado por 7 tablas cuyas descripciones se pueden ver en la Tabla 2. Para la generación del almacén de datos se tomaron e integraron datos de la base de datos de alumnos de la UNLaM, de la base de datos de encuestas del DIIT y de la bases de datos de colegios de educación secundaria del Ministerio de Educación.

Tabla 2. Descripción de las tablas.

Tablas	Descripción
Alumnos	Datos del estudiante.
Carreras	Datos de las carreras del DIIT.
PlanesEstudio	Datos de los planes de estudio, vigentes y no vigentes, de las carreras.
Materias	Datos de las materias de los planes de estudio.
Examenes	Datos de las notas, por carrera, plan de estudio y materia, de los estudiantes.
Censos	Datos de los censos realizados a los estudiantes.
Secundarios	Datos de los colegios de educación secundaria.

En la tarea de integración se transformaron los siguientes atributos:

- **fecha de nacimiento:** se redefinió este campo de tipo carácter con una longitud de 8 a tipo fecha con longitud determinada por la configuración del motor de base de datos.
- **año de ingreso:** se redefinió este campo de tipo carácter con una longitud de 2 a tipo numérico con una longitud de 4 sin decimales.
- **fecha de examen:** se redefinió este campo de tipo carácter con una longitud de 8 a tipo fecha con longitud determinada por la configuración del motor de base de datos.

3.2 Fase de limpieza, selección y transformación.

La calidad de los patrones que se obtienen con la minería de datos es directamente proporcional a la calidad de los datos utilizados. Esta fase es la responsable de obtener datos de alta calidad. Para lograr este objetivo se buscó detectar valores anómalos (outliers) y datos faltantes, se realizó una selección de los atributos relevantes y se construyeron nuevos atributos a partir de los existentes.

Para la detección de los valores anómalos y de los datos faltantes se realizó un análisis exploratorio. Del resultado de este análisis se desprendió que no había valores anómalos y que existían muy pocos datos faltantes, Se decidió reemplazar los datos faltantes por la moda del atributo en estudio.

Para validar la selección de los atributos relevantes realizada por el Secretario Académico del DIIT y los Coordinadores de las carreras de Informática, Electrónica e Industrial, se realizó un análisis de correspondencias cuyo resultado no modificó los atributos ya seleccionados.

Los atributos generados, para cada estudiante, fueron:

- **edad:** este atributo se generó a partir de la fecha de nacimiento.
- **indice_materias:** este atributo se generó tomado el resultado de la división de la cantidad de materias aprobadas por la cantidad de años entre la fecha actual o fecha de abandono y la fecha de ingreso. La cantidad de materias aprobadas se obtuvo de la cantidad de instancias en la tabla Examenes con un valor en el atributo Nota igual o mayor que 4. En la Tabla 3 se puede ver la discretización de este atributo.
- **reprobadas:** este atributo se generó a partir de la cantidad de instancias en la tabla Examenes con un valor en el atributo Nota menor que 4.
- **promedio:** este atributo es el cálculo del promedio del alumno.

Tabla 3. Discretización del atributo índice_materias.

índice_materias	Valor
Menor a 2	1 – Malo
Mayor a 1,99 y menor a 3	2 – Regular
Mayor a 2,99 y menor a 4,5	3 – Bueno
Mayor a 4,49 y menor a 5,5	4 – Muy bueno
Mayor a 5,49	5 – Excelente

3.3 Fase de minería de datos.

Dentro del proceso de DCDB está fase es la encargada de producir nuevo conocimiento [8]. En este trabajo se decidió utilizar:

- la clasificación como tipo de tarea de minería.
- el árbol de decisión como tipo de modelo.
- el J48 (implementación en Weka del algoritmo C4.5) [11] y el FT [5] como algoritmos de minería.

En la Tabla 4 se pueden ver los atributos del archivo elaborado para la fase de minería de datos. Este archivo contiene 9545 instancias que representan a los alumnos inactivos, activos y reincorporados. Para entrenar los modelos se utilizó un archivo

con 2865 instancias (30% del original), que fueron seleccionadas en forma aleatoria.

Tabla 4. Atributos del archivo utilizado en la fase de minería de datos.

Nombre	Descripción	Tipo
sexo	1 – Masculino 2 – Femenino	Nominal
edad	Edad	Numérico
estado_civil	1 – Casada/o 2 – Divorciada/o 3 – Soltera/o 4 – Separada/o 5 – Viuda/o	Nominal
carrera	201 – Ing. en Informática 202 – Ing. Electrónica 203 – Ing. Industrial	Nominal
estado	1 – inactivo 2 – activo 3 – reincorporado	Nominal
indice_materias	1 – Malo 2 – Regular 3 – Bueno 4 – Muy Bueno 5 – Excelente	Nominal
Promedio	Promedio del alumno	Numérico
reprobadas	Cantidad de materias no aprobadas	Numérico
trabajo	1 – No trabaja 2 – Trabaja	Nominal
horas	Total de horas trabajadas diariamente	Numérico
horario	1 – Mañana 2 – Tarde 3 – Noche	Nominal
gestion_escuela	1 – Estatal 2 – Privada	Nominal
tipo_escuela	1 – Bachiller 2 – Comercial 3 – Polimodal 4 – Técnica	Nominal
estudio_padre	1 – Sin Estudios 2 – Estudios primarios 3 – Estudios secundarios 4 – Estudios superiores	Nominal
estudio_madre	1 – Sin Estudios 2 – Estudios primarios 3 – Estudios secundarios 4 – Estudios superiores	Nominal

Se eligieron como clases los siguientes atributos:

- **indice_materias:** para encontrar los patrones determinantes del rendimiento académico.
- **estado:** para encontrar los patrones determinantes de la deserción estudiantil.

3.3.1 Rendimiento académico.

El mejor resultado fue obtenido por el algoritmo FT que alcanzó un 78,07% de instancias clasificadas correctamente, mientras que el algoritmo J48 clasificó en forma correcta un 72,53% de las instancias. En las Tabla 5 se puede observar la matriz de confusión generada por el algoritmo FT y en la Tabla 6 la generada por el algoritmo J48.

Tabla 5. Matriz de confusión generada por el algoritmo FT.

a	b	C	d	e	
3197	437	276	41	5	a = 1- Malo
184	2093	151	27	6	b = 2 - Regular
79	380	1357	34	13	c = 3 - Bueno
199	69	26	552	28	d = 4 - Muy bueno
16	54	23	45	253	e = 5 - Excelente

Tabla 6. Matriz de confusión generada por el algoritmo J48.

a	b	c	d	e	
3588	276	28	41	23	a = 1- Malo
387	1794	201	46	33	b = 2 – Regular
460	253	1081	52	17	c = 3 – Bueno
169	254	138	276	37	d = 4 - Muy bueno
32	60	69	46	184	e = 5 - Excelente

3.3.2 Deserción Estudiantil.

Al igual que en la clasificación anterior el mejor resultado fue obtenido por el algoritmo FT que alcanzó un 77,86% de instancias clasificadas correctamente contra el 72,78% logrado por el algoritmo J48. En las Tablas 7 y 8 se pueden ver las matrices de confusión generadas por los algoritmos FT y J48 respectivamente.

Tabla 7. Matriz de confusión generada por el algoritmo FT.

Inactivo	Activo	Reincorporado	
2344	716	183	Inactivo
390	3799	135	Activo
367	322	1289	Reincorporado

Tabla 8. Matriz de confusión generada por el algoritmo J48.

Inactivo	Activo	Reincorporado	
2435	694	114	Inactivo
687	3546	91	Activo
389	623	966	Reincorporado

3.4 Fase de evaluación e interpretación.

En un contexto ideal los patrones descubiertos por la fase de minería de datos deben reunir 3 cualidades: ser precisos, comprensibles e interesantes [8]. En este trabajo nos interesó mejorar principalmente la comprensibilidad.

Para efectuar la evaluación de los modelos se tomo como medida el porcentaje de aciertos al clasificar una instancia en su respectiva clase. Por cada algoritmo se realizaron 30 iteraciones

y en la Tablas 9 se pueden ver los mejores porcentajes de aciertos.

Tabla 9. Porcentaje de aciertos de los algoritmos de clasificación.

	FT	J48
Rendimiento académico	78.07%	72,53%
Deserción estudiantil	77,86%	72,78%

En la Tabla 9 se puede observar que el algoritmo FT tuvo un mejor desempeño que el algoritmo J48. Pero si se analizan las matrices de confusión (Tablas 5, 6,7 y 8) se puede ver que para detectar un rendimiento académico malo y alumnos inactivos el algoritmo J48 supera al FT (Tabla 10).

Tabla 10. Porcentaje de aciertos del rendimiento académico malo y de los alumnos inactivos.

	FT	J48
Rendimiento académico malo	80.81%	90,70%
Alumnos inactivos	72,28%	75,08%

Con respecto a la comprensibilidad de los modelos se puede decir:

- que el algoritmo J48 generó un árbol de decisión muy grande y por lo tanto poco comprensible y difícil de interpretar.
- que el árbol generado por el algoritmo FT no permite explicar el rendimiento académico y las causas de la deserción estudiantil.

3.5 Fase de difusión y uso.

Durante el curso del primer cuatrimestre se espera trabajar con un archivo que contenga los datos del año 2009 para poder seguir evaluando el poder predictivo de los modelos.

4 CONCLUSIONES Y TRABAJO FUTURO

El desarrollo de este trabajo permitió consolidar en el DIIT un grupo de investigación en las técnicas de Data Mining y además la implementación de un almacén de datos que permitirá tomar decisiones con menor incertidumbre.

Si bien no se logró encontrar un clasificador del rendimiento académico y de la deserción estudiantil con un alto grado de precisión y compresibilidad, se adquirió experiencia en el uso de los programas SPSS y Weka que permitirá que el grupo avance en esta línea de investigación.

Como trabajo futuro se buscará una solución para la predicción del rendimiento académico y la deserción estudiantil basada en la evolución de reglas. Se utilizará una red neuronal para la generación de las reglas y algoritmos genéticos para evolucionarlas.

5 REFERENCIAS

[1] Ballard C., Herreman D., Schau D., Bell R., Kim E., Valncic A.: "Data Modeling Techniques for Data Warehousing", IBM Red Book, 1998.

[2] Chen M., Han J., Yu P.: "Data Mining: An Overview from Database Perspective". IEEE Transactions on Knowledge and Data Engineering, 1996.

[3] Dapozo G., Porcel E., López M. V., Bogado V., Bargiele R.: "Aplicación de minería de datos con una herramienta de software libre en la evaluación del rendimiento académico de los alumnos de la carrera de Sistemas de la FACENA-UNNE". Anales del Octavo Workshop de Investigadores en Ciencias de la Computación WICC 2006. Morón, Buenos Aires, Argentina, 2006.

[4] Fayyad U. M., Piatetsky-Shapiro G., Smyth P.: "From Data Mining to Knowledge Discovery: An Overview". Advances in Knowledge Discovery and Data Mining pp:1-34, AAAI/MIT Press, 1996.

[5] Gama J.: "Functional Trees". Machine Learning pp:219-250, Springer Netherlands, 2004.

[6] Han, J., Kamber, M.: "Data Mining Concepts and Techniques". Morgan Kaufmann Publishers, San Francisco, 2006.

[7] Hand, D., Mannila, H., Smyth P.: "Principles of Data Mining". MIT Press, 2001.

[8] Hernandez, O. J., Ramirez, Q. M., Ferri, R. C.: "Introducción a la Minería de Datos". Editorial Pearson Prentice Hall, Madrid, España ,2004.

[9] Machine Learning Project at the Department of Computer Science of The University of Waikato, New Zealand. http://www.cs.waikato.ac.nz/ml/weka/

[10] Timarán, P. R.: "Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos". Memorias de la 8ª Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI 2009. Orlando, Florida, USA, 2009.

[11] Witten, I. H., Frank, E: "Data Mining Practical Machine Learning Tools and Techniques". Morgan Kaufmann Publishers, San Francisco, California, USA, 2005.