Resultados Preliminares del Proceso de Minería de Datos Aplicado al Análisis de la Deserción en Carreras de Informática Utilizando Herramientas Open Source

J. Germán A. Pautsch[1], Horacio D. Kuna[2], Antonia E. Godoy[3]
[1,2] Dpto. Informática. Facultad de Ciencias Exactas, Químicas y Naturales.
Universidad Nacional de Misiones
[3] Dpto. Matemática. Facultad de Ciencias Económicas.
Universidad Nacional de Misiones

(3300) Posadas. Argentina
[1] gpautsch@fceqyn.unam.edu.ar, [2] hdkuna@unam.edu.ar, [3] godoy@fce.unam.edu.ar

Resumen. En el presente trabajo se realizó un proceso de minería de datos para generar conocimiento en base a patrones académicos, factores sociales y demográficos, que caractericen a los estudiantes, con la finalidad de pronosticar alumnos desertores de la Carrera Analista en Sistemas de Computación de la Facultad de Ciencias Exactas, Químicas y Naturales de la Universidad Nacional de Misiones. Como fuente de datos de utilizó el "Cubo 04 Desgranamiento", exportado del Sistema de Gestión Académica SIU-Guaraní. Los modelos obtenidos se utilizaron para clasificar a los alumnos de otras cohortes. El trabajo se desarrolló bajo la metodología de libre difusión Crisp-DM y con herramientas open source. La calidad de los modelos obtenidos a través de la clasificación con árboles de decisión y redes bayesianas superaron ampliamente las expectativas.

Palabras Clave: Minería de Datos, Clasificación, Pronósticos, Deserción Universitaria, Perfiles de Alumnos.

1 Introducción

Se estima que las bases de datos (BD) de las organizaciones se duplican cada veinte (20) meses, según W.J. Frawley y otros [1]. Lamentablemente las técnicas de análisis de información no han tenido un desarrollo equivalente.

La Universidad Nacional de Misiones cuenta con el Sistema de Gestión Académica SIU-Guaraní (SIU-G) [2]. El sistema, produce una gran cantidad de datos, los cuales pueden ser muy valiosos, pero debido a su volumen resultan muy difíciles de analizar. Dentro de esta masa de datos hay información oculta de gran importancia que se podría llegar a descubrir con técnicas de minería de datos (MD).

Este artículo se ha estructurado de la siguiente manera: en la sección 2 se expone el objetivo principal del mismo, luego en la sección 3 se desarrolla una muy breve revisión de los principales conceptos de minería de datos, seguidamente en la sección 4 se indicarán los recursos disponibles y el software utilizado, para continuar en la sección 5 con la metodología seguida y en la sección 6 presentar algunos resultados obtenidos. Para finalizar en la sección 7 se brindan las conclusiones y trabajos futuros, y en la sección 8 las referencias consultadas.

2 Objetivo principal

El objetivo es realizar una MD, sobre las cohortes que se encuentran entre los años 2001 y 2006, a través de técnicas supervisadas, sobre el Cubo 04 exportado de la BD del SIU-G. De esta forma se busca determinar cuáles son las técnicas, algoritmos óptimos para extraer el conocimiento de la BD y así, confeccionar modelos para posteriormente intentar pronosticar con cierto grado de certeza, y en base a patrones académicos, factores sociales y demográficos, si un alumno posee o no características que aumenten su probabilidad de desertar de la carrera Analista en Sistemas de Computación.

La meta es lograr diseñar modelos cuya calidad de predicción o clasificación supere el 70%. Por otra parte se buscará estandarizar y automatizar los procesos E.T.L. (Extracción, Transformación y Carga de Datos) para que cada unidad académica pueda realizar la MD sobre la información exportada del SIU-G

3 Revisión conceptual

La MD se define formalmente como "*un conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir, de forma automatizada, tendencias o comportamientos y descubrir modelos previamente desconocidos*" [3].

Fundamentalmente, la diferencia de la MD con otras técnicas reside en que permite construir modelos de manera automática.

Cabe destacar que la MD es una etapa dentro de un proceso más amplio llamado Descubrimiento de Conocimiento en BD (Knowledge Discovery in Data Base – KDD).

En términos estrictamente académicos, los términos MD y KDD no deben utilizarse de manera indistinta. La MD es un paso esencial en el KDD que utiliza algoritmos para generar patrones a partir de los datos pre procesados [4] (Fig. 1).

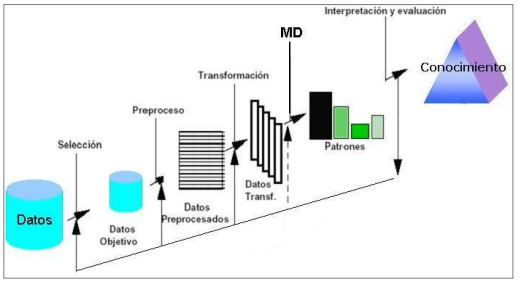

Fig. 1. Posicionamiento de la MD dentro de las etapas del KDD

El concepto de MD no es nuevo, pero han sido necesarios varios años de desarrollo para que esta técnica pudiera ser utilizada de manera sencilla.

La MD genera modelos que pueden ser descriptivos o predictivos [5].

- *Descriptivos o No Supervisados*: este modelo aspira a descubrir patrones y tendencias sobre el conjunto de datos sin tener ningún tipo de conocimiento previo de la situación a la cual se quiere llegar. Descubre patrones en los datos analizados. Proporciona información sobre las relaciones entre los mismos.
- *Predictivos o Supervisados*: crean un modelo de una situación donde las respuestas son conocidas y luego, lo aplica en otra situación de la cual se desconoce la respuesta. Conociendo y analizando un conjunto de datos, intentan predecir el valor de un atributo (Etiqueta), estableciendo relaciones entre ellos.

Uno de los factores claves que define la verdadera MD es que la aplicación misma realiza el análisis sobre los datos. En otros casos, el análisis es guiado por una interacción con el usuario. Las aplicaciones que no son, en algún grado, auto guiadas están realizando análisis de datos y no MD.

4 Recursos disponibles

4.1 Fuente de datos

Como se mencionó anteriormente, la Universidad Nacional de Misiones cuenta con el SIU-G. Realizando un relevamiento preliminar, se observó que en el mismo existe un módulo llamado "Interfaz" que exporta datos orientados al OLAP (*On Line Analytical Process* – Procesamiento Analítico En Línea) y abarcan diferentes temáticas. Luego de analizar detalladamente la documentación que describe cada opción de exportación [6], se determinó que el "Cubo 04 – Desgranamiento", puede ser de gran utilidad para

el presente trabajo, ya que aborda la temática de la deserción desde el punto de vista académico, social y demográfico.

Luego de aplicados los procesos de E.T.L., a continuación la Tabla 1 muestra la estructura y los nombres de los atributos de la Tabla DMS_C4 (Data Mining Source – Cubo 4), creada para el análisis del "Cubo 04 – Desgranamiento".

Tabla 1. Centralización de la información del "Cubo 4 – Desgranamiento" para el proceso de MD.

Atributos	Descripción	Valores
ACT_ANUAL	Actividad realizada durante el primer año académico Incluye cantidad de exámenes rendidos, promociones y equivalencias (no importa el resultado de los mismos). Es decir todo lo que refleje intención de aprobar una materia.	Sin actividad, 0<A<3, 2<A<6, A>5
SEXO	Género de la persona.	mujer, Varon
SIT_ESTUDIANTE	Se establecen dos categorías de estudiantes: los que no tuvieron actividad y los que tuvieron actividad con los últimos dos años	Activo (A), Pasivo (P)
ESTUDIO_PADRES	Indica el mayor nivel de estudios alcanzado por los padres del alumno.	No Posee, Pri., Sec., Uni.
LOCALIDAD	Localidad en la que se encuentra el colegio secundario del que egresó el estudiante	Posadas, Apostoles, Otras Loc.
COLEGIO	Orientación del colegio secundario del que egresó el estudiante	Comercial, Bachiller, Técnico, etc
PROVINCIA	Provincia a la que corresponde la anterior localidad.	Misiones, Correntes, Otras Prov.
DIST_SEDE	Distancia en kms a la sede donde se dicta la carrera desde la localidad de procedencia.	De 0 a 50 De 51 a 150 Etc.

4.2 Software utilizado

Las herramientas software utilizadas se enmarcas todas dentro de la filosofía open source, empleando para diseñar el modelo y realizar la MD RapidMiner v5.0 [7], para los procesos E.T.L. Pentaho Data Integration v4.1.0(PDI) [8] y para la crear la Almacén de Datos MySQL v5.0. Este ultimo recomendado por la Suite Pentaho.

5 Metodología

La metodología seleccionada fue CRISP-DM [9], ya que esta abarca una perspectiva más amplia contemplando también los objetivos empresariales del proyecto. [10].

Otras metodologías como SEMMA [11], está ligada a los productos de SAS Institute donde se encuentra implementada. La metodología CRISP-DM es una metodología libre y gratuita que no depende de la herramienta que se utilice para el desarrollo del proyecto de data mining.

La metodología CRISP-DM se organiza en seis etapas. Cada una de ellas a su vez se divide en varias tareas (Fig. 2), las flechas muestran las relaciones más habituales entre las etapas, aunque se debe aclarar que pueden establecer relaciones entre cualquiera de las fases. El círculo exterior ilustra la naturaleza cíclica del proceso de modelado.

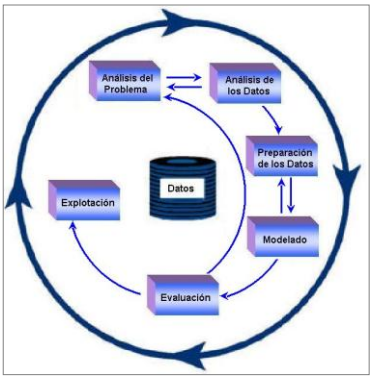

Fig. 2. Fases del proceso de modelado de la metodología CRISP-DM.

6 Resultados Obtenidos

6.1 Automatización de la importación de datos

Esta tarea fue llevada a cabo con la herramienta *Pentaho Data Integration* (PDI).

Se observa que por cada dimensión y tabla de hechos del "Cubo 04 Desgranamiento", el sistema SIU-G genera un archivo de texto. Entonces se creó un almacén de datos con una estructura similar a la exportada, con la finalidad de persistir los datos y así poder trabajar sobre ellos de manera óptima.

Luego se procede con la carga de la BD a través de la generación de "transformaciones". Las mismas se encargan de leer los archivos de texto e importarlos a una tabla destino correspondiente. Se debe definir una transformación

para cargar cada tabla de la BD. En esta instancia es donde se hacen correlacionar las columnas de la fuente de datos, archivos txt, y su correspondiente atributo en la tabla destino.

Posteriormente se define un "trabajo". El objetivo del mismo es integrar y ordenar la ejecución de cada transformación para controlar y optimizar la carga de la BD.

En la Fig. 3 se aprecia un flujo de trabajo el cual se inicia cargando la tabla "países". En el caso que la carga sea exitosa se continúa con la tabla "provincias" y así sucesivamente con las demás. De producirse un error, el mismo se escribe en un *log* y luego se aborta la ejecución. Si la carga de todas las tablas de dimensiones es exitosa, por ultimo se procede a cargar la tabla de hechos "desgranamiento".

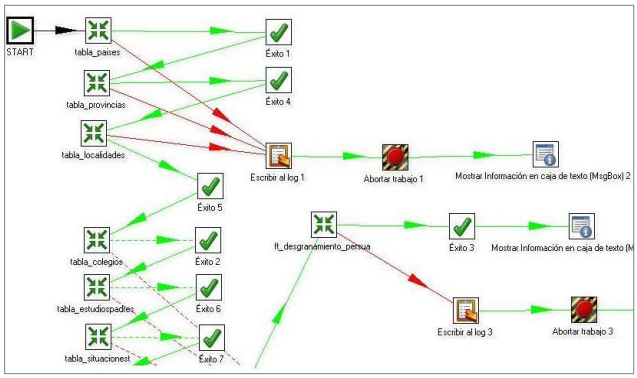

Fig 3. Trabajo definido para la carga del Cubo 4 - Desgranamiento

6.2 Minería de datos

Para llevar a cabo la MD se utilizo la herramienta *Rapid Miner.*

En esta sección presentamos los mejores resultados, obtenidos al ejecutar los flujo de minería con el componente *Decisión Tree*, que es una implementación del algoritmo C4.5 o también llamado CART.

Respecto a la clasificación obtenida, algunas de las reglas que el algoritmo ha podido establecer para la clasificación entre alumnos activos y pasivos se pueden observar a continuación en la Fig.4.

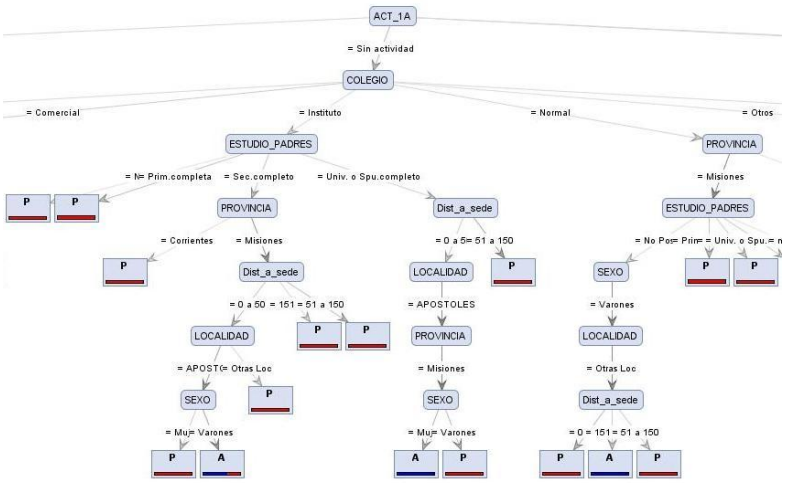

Fig 4. Árbol de Decisión obtenido con la herramienta Rapid Miner.

Posteriormente se realizo la validación del modelo, sometiéndolo a la clasificación con datos reales y previamente desconocidos.

En referencia a la cantidad de datos definida para el entrenamiento del algoritmo, podemos decir que la proporción de error disminuye a medida que la cantidad de datos de entrenamiento aumenta.

El flujo de minería para realizar la validación del modelo, puede observarse en la Fig.5.

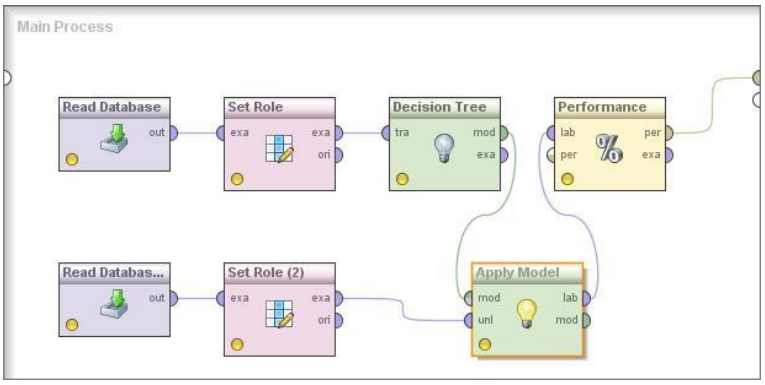

Fig 5. Flujo de minería para la validación del modelo obtenido.

En la parte superior de flujo el componente *Read Database* lee de la BD la cohortes desde el 2001 al 2006, luego el componente *Set Role* indica que los valores del atributo "Sit_Estudiante" va a ser la etiqueta (*label*) a predecir por el algoritmo

Decision Tree. Paralelamente en la parte inferior del flujo se ingresan los datos a pronosticar, ellos son los pertenecientes a la cohorte 2000. Con el componente *Aply Model* aplicamos el modelo obtenido a los datos desconocidos previamente para luego, con el componente *Performance*, podamos visualizar el rendimiento a cerca de la clasificación realizada sobre los datos.

En la matriz de confusión de la Fig. 6 podemos observar que el modelo clasificó incorrectamente sólo a siete (7) alumnos sobre un total de ciento noventa y seis (196) alumnos alcanzando una precisión del noventa y seis por ciento (96%)

accuracy: 96%

	true P	true A	class precision
pred. P	96	7	93%
pred. A	0	93	100%
class recall	100%	93%	

Fig. 6. Matriz de confusión resultando de la validación del modelo

7 Conclusiones y trabajos futuros

En cuanto a la interpretación de los resultados, esta se delego a los expertos en el dominio de la deserción. Todos ellos han observado que, si bien se realiza una muy buena clasificación de los alumnos Activos y Pasivos, salvo el Nivel de Estudio de los Padres, la localidad, el desarraigo (atributo Dist_a_Sede) y el colegio, no existen otras variables relevantes al análisis socio económico de la deserción estudiantil. Sería interesante poder incorporar al estudio, indicadores que permitan saber si el alumno tiene personas a cargo, si trabaja, si es que viaja para cursar, etc.

Como conclusiones del lado del ingeniero en conocimiento, primeramente se debe comentar que en este trabajo sólo se han abarcado algunos métodos de extracción del conocimiento a través de la MD. No obstante, existen muchas más posibilidades que ofrecen ésta y otras herramientas.

Queda demostrado que para realizar una minería de datos de buena calidad, ésta debe estar acompañada de una serie de mecanismos, transformaciones, flujos de trabajo, modelos de validación, matrices de confusión, etc., que faciliten y permiten realizar una validación y un análisis de resultados más completo y fiable.

Con la aplicación de árboles de decisión y redes bayesianas se han obtenido muy buenos resultados, superando lo planteado como objetivo específico de la MD. La aplicación de cada algoritmo facilitó advertir, no sólo las diferentes características pertenecientes al grupo de alumnos Pasivos, sino que también han quedado manifestadas las características de las clases contrastes (alumnos Activos y Egresados).

Las redes bayesianas permitieron advertir mas detalladamente cuáles eran los atributos más importantes por el cual el algoritmo realizaba la clasificación de los alumnos.

Como contrapartida, la interpretación del Árbol de Decisión obtenido, puede resultar difícil de leer por personas no especializadas, debido a su amplitud. Esta dificultad es compensada por la muy buena representación gráfica que implementa la herramienta y la posibilidad de exportar las regles de decisión.

Si bien la calidad de los modelos superó las expectativas planteadas, se considera muy importante contar con la opinión de los expertos, no sólo a la hora de crear los modelos sino que también en lo que refiere a la evaluación e interpretación de los resultados

Un aporte muy significativo es el haber logrado automatizar los procesos ETL a través de la implementación de transformaciones y flujos de trabajo. Con PDI a su disposición, la Unidad Académica, que así lo desee, podrá extraer el conocimiento de sus BD con más facilidad evitando largas etapas de pre proceso.

Dada la flexibilidad que otorga la herramienta, no representaría mayor inconveniente, el introducir más variables socio económicas, como sugieren los expertos.

A lo largo del desarrollo del presente trabajo han surgido varias líneas para ser abordadas en el futuro.

Entre algunas de ellas podemos mencionar:

- Incorporar otras fuentes de datos que contengan más variables socio económicas como estado civil, situación laboral, familiares a cargo y otras contenidas en la BD del SIU-G, particularmente en la tabla sga_Datos_Censales, y las sugeridas por los expertos
- Diseñar nuevos flujos de minería incorporando otros algoritmos como los referentes a clusterización, regresión, correlación, etc.
- Implementar un tablero de control sobre las variables más relevantes detectadas en el proceso de minería
- Establecer alguna métrica para medir la información y la confusión que aporta cada atributos en referencia a la variable a predecir

Agradecimientos. A mis tutores por brindarme su tiempo y conocimientos para el desarrollo del presente trabajo. A todos mis alumnos, especialmente a Martín Rey y Cinthia Cuba. A mi familia, por regalarme sus sonrisas y ternura a pesar de mis ausencias.

8 Referencias

1. Frawley, W.J.; Piatetski-Shapiro, G.; Matheus, C.J. "*Knowledge Discovery in Databases*", AAAI-MIT Press (1991)

2. Consorcio SIU. Ministerio de Educación Ciencia y Tecnología. Secretaria de Políticas Universitarias. http://www.siu.edu.ar/acerca_de/que_es_el_siu. Accedido el 13 de Septiembre de 2009

3. Frawley, W.J.; Piatetski-Shapiro, G.; Matheus, C.J. *"Knowledge Discovery in Databases: an Overview"*. AI Magazine (1992)

4. Frawley, W.J.; Piatetski-Shapiro, G.; Smyth, P., "*From Data Mining to Knowledge Discovery in Databases*". AAAI-MIT Press (1996)

5. Agrawal, R.; Shafer, J. C. *"Parallel Mining of Association Rules"* IEEE Transactions on Knowledge and Data Engineering. (1996)

6. Consorcio SIU. Ministerio de Educación Ciencia y Tecnología. Secretaria de Políticas Universitarias *"Descripción del Cubo 04 Desgranamiento".* (2005)

7 Rapid-I GmbH. *"Rapid Miner 5.0 User Manual".* Dortmund (2010)

8 Roldán, M. C. *"Primeros pasos con Pentaho Data Integration"*. (2009)

9. Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.;Wirth, R. *"CRISP-DM 1.0. Step-by-step data mining guide".* (1999)

10. Gondar, J. E. *"Comparación de Metodologías de Data Mining".* Accedido el 23 de Julio de 2009

11. Sas Institute. http://www.sas.com/technologies/analytics/datamining/miner/semma.html. Accedido el 20 de Junio de 2009