Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2005 Proceedings

Americas Conference on Information Systems (AMCIS)

2005

Using Data Mining to Support the University Decision Process: A Case in a Chilean University

Aurora D. Sanchez
Universidad Catolica del Norte, asanchez@ucn.cl

Marco A. Gutierrez
Universidad Catolica del Norte - Chile, marcogut@ucn.cl

Claudio V. Meneses
Universidad Catolica del Norte - Chile, cmeneses@ucn.cl

Using Data Mining to Support the University Decision Process: A Case in a Chilean University

Aurora D. Sánchez
Universidad Catolica del Norte
asanchez@ucn.cl

Marco A. Gutiérrez
Universidad Catolica del Norte
marcogut@ucn.cl

Claudio V. Meneses
Universidad Catolica del Norte
cmeneses@ucn.cl

ABSTRACT
Data mining is increasingly becoming an essential tool in organizations today. Particularly, academic organizations are requesting more sophisticated tools to improve their decision making process. A large quantity of data and information is produced during the student's life, but it is still necessary to turn them into insight. This paper describes a project that use data mining to support the decision making process in higher education in Chile. The aim of this project is to find patterns that allow the identification and determination of relationships among the initial conditions of students and with their final status as a student (drop-out or graduated). The study is conducted in a university in the north of Chile and it considers five undergraduate majors. The final results of this project are expected to support the decisions making process related with university admission policies, causes of student failure or success, and university marketing policies.

Keywords

Data mining, Decision Support, Educational Data Analysis.

INTRODUCCIÓN

Debido a la evolución de las tecnologías de la información en la última década y a la automatización de la mayoría de los procesos de negocios, se ha producido un aumento en las bases de datos empresariales. Gran parte de estos datos son históricos, es decir, que representan transacciones y situaciones ocurridas en el pasado.

La utilización de estos datos es útil para tomar medidas sobre eventos futuros ya que la mayoría de las decisiones se toman apoyándose en la experiencia pasada. El problema que se produce es que el análisis y aprovechamiento de la información que allí se encuentra se ha vuelto cada vez más complejo. Por este motivo es que surge el "Descubrimiento de Conocimiento a partir de Bases de Datos" (KDD, del ingles Knowledge Discovery from DataBases), definido como el "proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en ultima instancia comprensibles a partir de los datos" (Fayyad, Piatesky-Shapiro, and Smyth, 1996). Dentro de este contexto aparece el término Data Mining (Minería de Datos). El proceso del KDD esta detallado en la Figura 1.

En el ámbito de la educación superior, se almacenan muchos datos de los estudiantes que ingresan, permanecen y egresan de ellas. Sin embargo, gran parte de estos datos no se utilizan para generar modelos que aporten conocimientos para la toma de decisiones. Este artículo presenta un caso de estudio de la aplicación del proceso y técnicas de Data Mining a datos de alumnos de una universidad chilena. El proyecto persigue encontrar patrones que permitan relacionar las condiciones iniciales de los estudiantes con su condición final, es decir graduado o eliminado, induciendo modelos y validando la hipótesis con cuatro carreras de pre-grado: Arquitectura, Ingeniería en Computación, Derecho, Geología, y Periodismo. Los resultados obtenidos permitirán soportar decisiones respecto a políticas de admisión, campañas de difusión, y optimizar recursos enfocados a mejorar las tasas de retención y perfiles de ingreso de los estudiantes de la universidad bajo estudio.

CONCEPTOS PRELIMINARES

El nombre Data Mining deriva de las similitudes entre buscar información valiosa en grandes bases de datos y minar una montaña para encontrar una veta de materiales valiosos. Ambos procesos requieren examinar una inmensa cantidad de material hasta encontrar exactamente donde residen los valores.

Data Mining, es definido como *"un proceso de descubrimiento de nuevas y significativas correlaciones, patrones y tendencias, examinando grandes cantidades de datos almacenados y usando técnicas de reconocimiento de patrones, de estadísticas y matemáticas"* (Gartner Group, 2004).

Data Mining puede ser visto de dos formas:

- ***Data Mining como una etapa dentro del proceso KDD***. (Fayyad y otros, 1996) Como se muestra en la Figura 1.

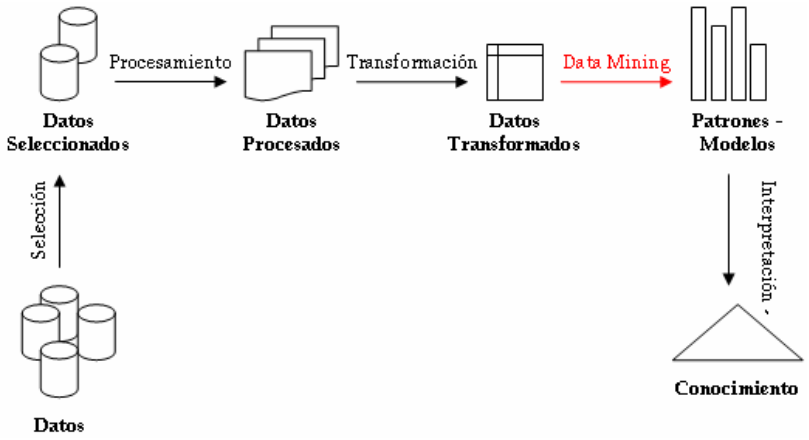

Figura 1. Etapas del KDD (Fayyad y otros, 1996)

- ***Data Mining como un Proceso independiente*** que posee sus propias etapas. (Chapman y otros, 2000)

Técnicas de Data Mining

La aplicación de las técnicas de Data Mining, tiene dos fines fundamentales: Construir de modelos y detectar patrones (Hand , Mannila, and Smyth, 2001). La construcción de modelos busca producir un resumen del conjunto de datos para identificar y describir las principales características. La detección de patrones busca identificar pequeñas desviaciones de la norma, para detectar patrones inusuales de comportamiento a través del descubrimiento de Patrones y Reglas y de Búsquedas por contenidos. Cuando no es posible construir modelos para el conjunto de datos, se pueden buscar patrones de comportamiento. El descubrimiento de patrones y reglas busca encontrar combinaciones y/o asociaciones de atributos que ocurren con frecuencia en transacciones de bases de datos (por ejemplo productos que se adquieren juntos). Este problema se ha atacado mediante el uso de técnicas basadas en reglas de asociación.

Áreas de Aplicación

Data Mining ha sido aplicada a una variedad de dominios e industrias, algunas de las cuales son:

- Marketing y Análisis de Canasta de Productos. La aplicación más conocida de data mining tiene relación con campañas de marketing de productos dirigidos a segmentos de clientes que optimicen los resultados de éstas.
- Análisis de las bases de datos comerciales para detectar terroristas. Por ejemplo, en julio de 2002, el director del FBI, John Aschcroft, anunció que el Departamento de Justicia se introduciría en los datos comerciales de hábitos y preferencias de compra de los consumidores, con el fin de descubrir potenciales terroristas antes de que ejecuten una acción.
- La detección de fraudes con tarjetas de crédito es también otro área de aplicación de data mining. En el año 2001, las instituciones financieras a escala mundial perdieron más de 2.000 millones de dólares en fraudes con tarjetas de crédito y débito.
- Predicción del tamaño de las audiencias televisivas. La British Broadcasting Corporation (BBC) utiliza un sistema para predecir el tamaño de las audiencias televisivas para un programa propuesto, así como el tiempo óptimo de exhibición (Brachman y otros, 1996). El sistema utiliza redes neuronales y árboles de decisión.

- En la educación superior en México, en un estudio de los egresados de la universidad en la carrera de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Chihuahua II, se uso Data Mining para observar si los graduados de la carrera se insertaban en actividades profesionales relacionadas con sus estudios.

Metodología De Desarrollo De Proyectos De Data Mining

El desarrollo de un proyecto usando data mining debe ser estructurado por lo cual es necesario contar con una metodología de desarrollo. La metodología CRISP-DM (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, and Wirth, 2000) es una de las más eficientes. En esta metodología, el proceso está organizado en seis fases. Las tareas generales se proyectan a tareas específicas, donde se describen las acciones que deben ser desarrolladas para situaciones específicas, como aparece en la Figura 2.

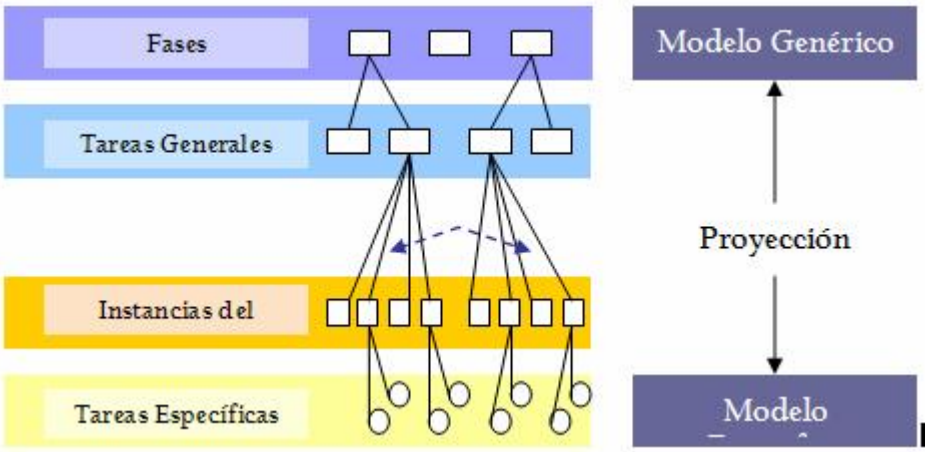

Figura 2. Tareas en la Metodología CRISP-DM (Chapman y otros, 2000)

El objetivo del método es permitir a diferentes empresas usar el mismo vocabulario, metodología y herramientas en las actividades de DM. Las 6 etapas de la metodología se presentan en la Figura 3.

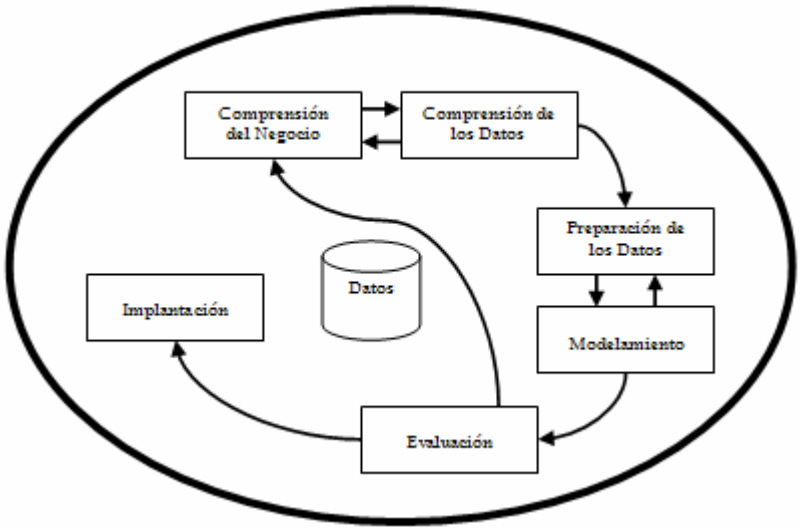

Figura 3. Etapas de la Metodología CRISP-DM (Chapman y otros, 2000)

Comprensión del Negocio. La primera fase análisis del problema, incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial o institucional.

Comprensión de los Datos. La segunda fase de análisis de datos comprende la recolección inicial de datos, identificando la calidad de los datos.

Preparación de los Datos. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Modelamiento. En la fase de modelado se seleccionan las técnicas de modelado más apropiadas para el proyecto. Una vez seleccionada las técnicas se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos.

Evaluación. En la fase de evaluación, se evalúa el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del problema. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la explotación del modelo.

Implantación. En esta etapa, además de la implantación del modelo, los resultados deben presentarse y documentarse de forma comprensible, para lograr un incremento del conocimiento.

EL CASO DE LA UNIVERSIDAD CATÓLICA DEL NORTE

El presente estudio emplea la metodología de análisis de caso apoyada por Data Mining para conocer de que manera las condiciones iniciales del alumno, al ingresar a la Universidad Católica del Norte (UCN), permiten entender su éxito o fracaso académico. Este estudio se realizo para el caso del proceso de admisión de la UCN. Las etapas del estudio se desarrollaron a partir de CRISP-DM y se presenta en la Figura 4.

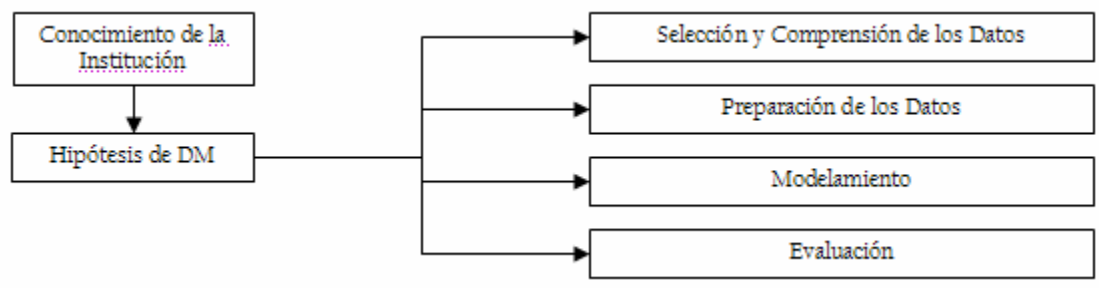

Figura 4. Estructura del proyecto Data Mining en la UCN.

Conocimiento de la Institución y Definición de la Hipotesis

La Universidad Católica del Norte (UCN) es una institución de educación superior que fue fundada en el 31 de Mayo de 1956 y está ubicada en las regiones II y IV de Chile.

Uno de los mayores problemas que enfrenta la universidad son las altas tasas de deserción estudiantil (Mujica, 2004). La Tabla 1 presenta los porcentajes de alumnos eliminados, egresados y titulados de las carreras en análisis para las cohortes de entre los años 1995 y 2004 de la UCN. Es por estas razones que el objetivo general del proyecto fue descubrir si existen patrones que ayuden a determinar las condiciones iniciales que expliquen la eliminación, egreso o titulación.

Carrera/Universidad	Eliminados (%)	Egresados (%)	Titulados (%)
Universidad en General	37.4	6.33	9.7
Arquitectura	40.56	2.14	4.01
Ingeniería Ejecución en Computación	49.46	1.53	7.35
Derecho	37.73	3.08	18.04
Periodismo	24.25	8.53	23.95
Geología	34.54	5.45	0.90

Tabla 1. Porcentajes de eliminación, egreso y titulación carreras UCN.

La hipótesis que se plantea en este estudio es:

H1: Las condiciones iniciales de notas de enseñanza media, prueba nacional de ingreso a la universidad (PSU) y colegio de origen permiten predecir si un estudiante finalizará exitosamente o no su carrera.

En base a conocimiento de experto del dominio del problema, se seleccionaron cuatro variables de entrada fundamentales en la formación escolar del alumno: promedio de notas de enseñanza media, puntaje de selección sin bonificación con el que ingreso a su carrera, tipo de colegio del que proviene, tipo de enseñanza que se impartía en su colegio. Las carreras que se incluyeron en el análisis para la validación de las hipótesis fueron Arquitectura, Ingeniería Ejecución en Computación, Derecho, Periodismo, Geología.

Selección y Comprensión de los Datos

Para la comprobación de la hipótesis fue necesario disponer de los datos del proceso de admisión y las bases de datos de los alumnos eliminados y los egresados. Este tipo datos se localizan en dos grandes fuentes: Data Warehouse de la UCN y las tablas de la base de datos del sistema curricular "SIMBAD" de la Universidad.

Del Data Warehouse de la UCN se utilizó el Data Mart de Admisión y de la base de datos del SIMBAD se seleccionaron las tablas PLANES_ALUMNOS, TIPO_PLANES_ALUMNOS y SEMESTRE, para obtener los datos del último estado académico que tiene (o que alcanzó) un alumno. Este estado puede ser "regular", "egresado", "eliminado", "titulado".

Para la creación de los modelos no se trabajo directamente con la base de datos, sino que se exportaron las tablas a archivos de textos separados por tabuladores como aparece en la Figura 5.

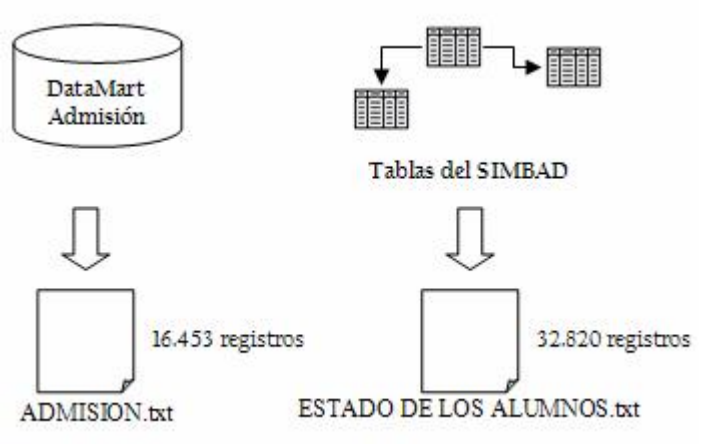

Figura 5. Conversión de las Fuentes de Datos

Preparación de los datos

Los cambios necesarios a los archivos se realizaron con la herramienta Clementine 8.1debido a sus capacidades de presentación que hacen más comprensible el resultado al usuario final.

1. **Selección de los atributos y limpieza de los registros:** primero sen seleccionaron los atributos que se utilizarían teniendo en cuenta los objetivos planteados. Para cada tabla se seleccionaron los siguientes atributos:

   ADMISION : (*año de proceso admisión, código de la carrera, nombre de la carrera, numero de inscripción, identificador persona, rango puntaje de selección sin bonificación, rango nota de enseñanza media, RUT del alumno, semestre proceso de admisión, sistema de admisión, tipo de colegio, tipo de educación en colegio, sede de la carrera*).

   ESTADO DE LOS ALUMNOS**:** (*semestre de ingreso, año de proceso admisión, estado del alumno, identificador persona, RUT del alumno*).

   Para el análisis se plantearon los siguientes filtros:

- Alumnos que ingresaron a la Universidad entre los años 1995 y 2004: (*año de proceso admisión ≥ 1995)* and *(año de proceso admisión ≤ 2004).*
- Alumnos que ingresaron de manera regular: *sistema de admisión* = "regular"
- Alumnos que ingresaron a carreras que tienen alumnos egresados y que no tienen un plan común.
- Alumnos que se encuentran egresados o eliminados de la Universidad. *estado del alumno*= "egresado" or "egresado eliminado" or "titulado" or "eliminado"

2. **Calidad de los Datos:** un problema que presentan los datos de admisión, es la cantidad de datos faltantes en algunos de los atributos *tipo de colegio, tipo de educación en colegio, rango puntaje de selección sin bonificacion, rango nota de enseñanza media.* Se decidió conservar los registros con valores desconocido debido a que al eliminar, por ejemplo los registros con *tipo de colegio* ="Desconocido" se excluían también filas con valores válidos para el *rango puntaje de selección sin bonificacion* y *rango nota de enseñanza media* .

3. **Construcción de los Datos**: se creó un nuevo atributo llamado *lograegresar*, que deriva del atributo *estado del alumno*. *lograegresar* toma el valor "si" cuando *estado del alumno* es "egresado" o "titulado" o "egresado eliminado" y es "no" cuando *estado del alumno* es "eliminado"

Cuando se terminaron los procesos de selección y construcción de datos, se guardaron los cambios realizados a los archivos para que después pudiesen utilizarse en la etapa de modelamiento. Los nuevos archivos quedaron en formato SPSS como aparece en la Figura 6.

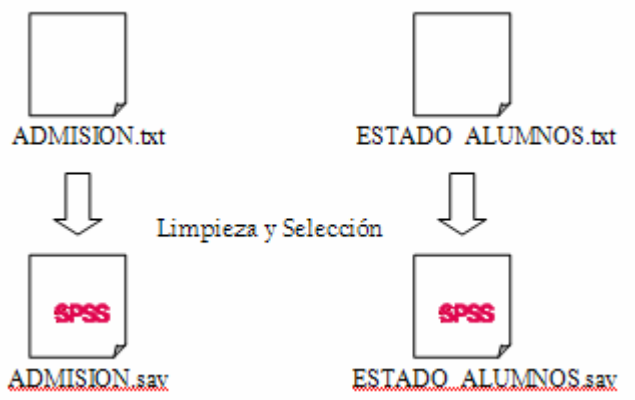

Figura 6. Nuevos archivos

El sistema de calificaciones en Chile va desde 1 a 7 donde la nota de aprobación es 4. La UCN recibe alumnos con puntajes en la prueba nacional de admisión PSU de mínimo 450 puntos (el puntaje máximo en la PSU es 800 puntos) La UCN bonifica con un puntaje adicional de 30, 20 ó 10 puntos a las tres primeras preferencias a sus carreras. Los alumnos deben tener un puntaje de selección sin bonificación de mínimo 500 puntos.

Modelamiento

El modelamiento de los datos se realizó para las la UCN a nivel global, para la UCN considerando carreras, y para las cinco carreras mencionadas. Los resultados del modelamiento en extenso se presentan a continuación para los casos generales de la UCN, el caso de las carreras de Arquitectura y el caso de Derecho. Dichas carreras se seleccionaron por sus particulares diferencias.

CASO 1: Análisis Global para la UCN.

El primer caso consideró el análisis todos los datos que se tenían para verificar si las condiciones de formación escolar del alumno pueden predecir su egreso de la universidad.

Como se quería trabajar con modelos predictores o clasificadores se optó por trabajar con árboles de decisión y redes neuronales. Para este tipo de modelos se debe contar con un objetivo (variable a predecir o variable dependiente) y un conjunto de entradas al modelo (variables independientes). Como el objetivo es predecir el egreso la variable dependiente será LOGRAEGRESAR. Las variables independientes serán *rango nota de enseñanza media, rango puntaje de selección sin bonificacion, tipo de educacion en colegio, tipo de colegio*. La Figura 7 presenta el modelo construido con Clementine 8.1.

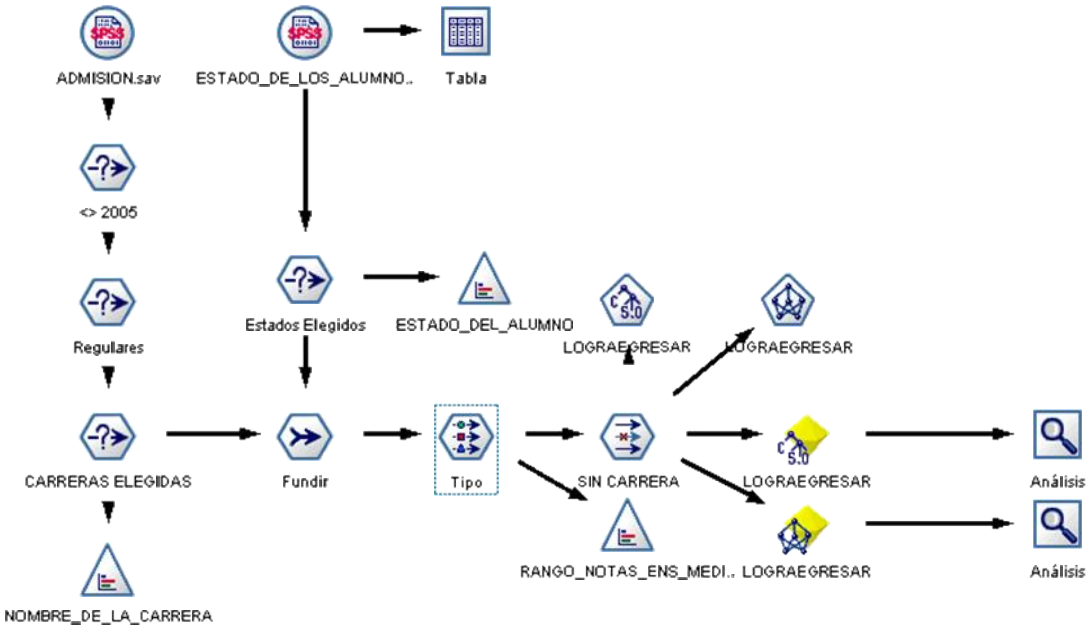

Figura 7. Modelo en Clementine 8.1 para el CASO 1

El resultado obtenido del modelamiento con redes neuronales es el siguiente (Figura 8):

Resultados para el campo de resultado LOGRAEGRESAR

Comparando $N-LOGRAEGRESAR con LOGRAEGRESAR

Correctos	3394	70,81%
Erróneos	1399	29,19%
Total	4793	

Matriz de coincidencias para $N-LOGRAEGRESAR (las filas muestran las reales)

	NO	**SI**
NO	3228	86
SI	1313	166

Figura 8. Resultados para análisis Global UCN redes neuronales

La red clasifica correctamente al 70,81% de los registros. El problema es que ésta, junto con la matriz de coincidencias, es la única información que aporta. No es posible saber como realizó la clasificación ni que atributos tomó como importantes.

Esto cambia al utilizar el algoritmo de árboles de decisión. Los resultados de la clasificación son los siguientes (Figura 9):

Resultados para el campo de resultado LOGRAEGRESAR

Comparando $C-LOGRAEGRESAR con LOGRAEGRESAR

Correctos	3402	70,98%
Erróneos	1391	29,02%
Total	4793	

Matriz de coincidencias para $C-LOGRAEGRESAR (las filas muestran las reales)

	NO	**SI**
NO	3098	216
SI	1175	304

Figura 9. Resultados para análisis Global UCN con árboles de decisión

Además de esto, es posible observar la estructura del árbol y así tomar decisiones sobre los atributos que fueron usados. Algunas conclusiones que pueden obtenerse del modelo de árboles de decisión son que la variable más importante para

predecir el egreso es la nota de enseñanza media y que los alumnos con nota enseñanza media menor que 6,0 son clasificados por el árbol como que NO logran egresar.

Si bien el modelo tiene un alto porcentaje de certeza (70,98%). Su estructura no arroja mucha información, por esta razón, se realizó un más análisis particular.

CASO 2: Análisis Particular para la UCN

Este caso es una repetición del anterior, con la diferencia es que ahora se agrega como variable independiente (entrada al modelo) el atributo *nombre de la carrera*. La porción que se muestra en la Figura 10 son los nodos agregados al modelo anterior.

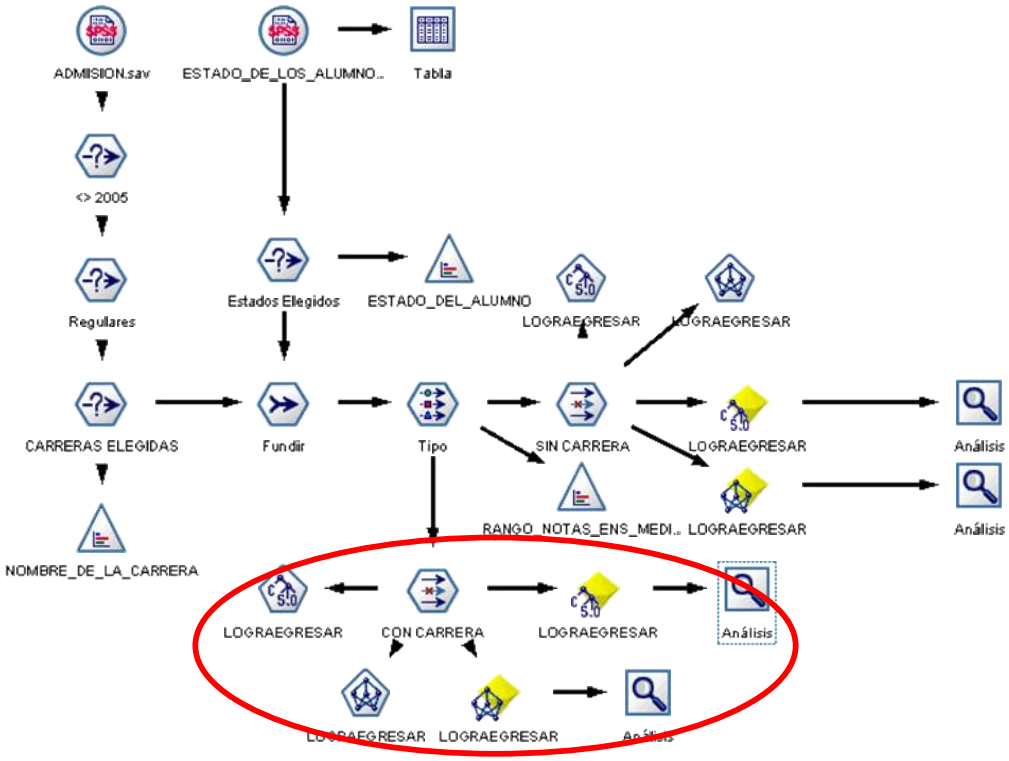

Figura 10. Modelo en Clementine 8.1 para el caso 2

El modelo construido con redes neuronales da el siguiente resultado (Figura 11):

Resultados para el campo de resultado LOGRAEGRESAR

Comparando $N-LOGRAEGRESAR con LOGRAEGRESAR

Correctos	3568	74,44%
Erróneos	1225	25,56%
Total	4793	

Matriz de coincidencias para $N-LOGRAEGRESAR (las filas muestran las reales)

	NO	SI
NO	2847	467
SI	758	721

Figura 11. Resultados análisis para caso particular UCN redes neuronales

El árbol de decisión entrega el siguiente resultado (Figura 12):

Resultados para el campo de resultado LOGRAEGRESAR

Comparando $C-LOGRAEGRESAR con LOGRAEGRESAR

Correctos	3543	73,92%
Erróneos	1250	26,08%
Total	4793	

Matriz de coincidencias para $C-LOGRAEGRESAR (las filas muestran las reales)

	NO	**SI**
NO	2957	357
SI	893	586

Figura 12. Resultados para análisis particular UCN con árboles de decisión

El árbol clasifica correctamente al 73,92% de los registros, sin embargo, él mayor conocimiento que entrega este modelo es que no se pueden tomar a todos los alumnos de la Universidad y analizarlos de la misma manera, ya que la variable carrera es la más importante para predecir el egreso.

Un factor importante es que para algunas carreras, como por ejemplo Arquitectura, Geología y Derecho, el árbol clasifica a sus alumnos como no egresados.

Cada carrera es clasificada de diferente forma por el árbol por lo tanto el siguiente paso es analizar los atributos de formación escolar para algunas carreras de la Universidad. Las carreras escogidas para este análisis son: Arquitectura, Ingeniería Ejecución en Computación e Informática, Derecho, Geología, Periodismo.

CASO 3: Análisis de Carrera de Arquitectura

La carrera de arquitectura es una de las carreras con mayor tasa de eliminación en la UCN. Además, es una de las carreras que no era clasificada por el árbol construido en el caso 2.

Se utilizó un algoritmo de reglas de asociación ver si es posible identificar algunos patrones interesantes y útiles. La configuración del algoritmo es la siguiente (Figura 13).

Como resultado se obtienen 39 reglas de asociación para el atributo estado del alumno = “eliminado”.

Al inspeccionar las reglas es fácil darse cuenta de que no hay un patrón para la carrera de arquitectura, debido a que todos los valores posibles de los atributos aparecen en ellas.

- Análisis
  - Profundidad del árbol: 1
- Campos
  - Objetivo
    - ESTADO_DEL_ALUMNO
  - Entradas
    - TIPO_DE_EDUCACION_EN_COLEGIO
    - TIPO_DE_COLEGIO
    - RANGO_NOTAS_ENS_MEDIA
    - RANGO_DEL_PUNTAJE_SB
- Configuración de creación
  - Número máximo de antecedentes: 5
  - Soporte mínimo de las reglas: 10,0
  - Confianza mínima de las reglas: 80,0
  - Optimizar: Velocidad
  - Sólo valores verdaderos para las marcas: verdadero
  - Datos transaccionales: falso
- Resumen de entrenamiento

Figura 13. Resultados para caso arquitectura con reglas de asociación

CASO 4: Análisis de Carrera de Derecho

El caso de derecho fue analizado utilizando los 3 tipos de algoritmos ya vistos: redes neuronales, árboles de decisión y reglas de asociación. Además se agrego el atributo *sede de la carrera*, para verificar si hay alguna distinción entre Derecho – Antofagasta y Derecho – Coquimbo. El resultado que arroja la red neuronal es el siguiente (Figura 14):

Resultados para el campo de resultado LOGRAEGRESAR

Comparando $N-LOGRAEGRESAR con LOGRAEGRESAR

Correctos	477	70,56%
Erróneos	199	29,44%
Total	676	

Matriz de coincidencias para $N-LOGRAEGRESAR (las filas muestran las reales)

	NO	SI
NO	399	57
SI	142	78

Figura 14. Resultados para caso Derecho con redes neuronales

Y el árbol de decisión (Figura 15):

Resultados para el campo de resultado LOGRAEGRESAR

Comparando $C-LOGRAEGRESAR con LOGRAEGRESAR

Correctos	487	72,04%
Erróneos	189	27,96%
Total	676	

Matriz de coincidencias para $C-LOGRAEGRESAR (las filas muestran las reales)

	NO	SI
NO	422	34
SI	155	65

Figura 15. Resultados para caso Derecho con árboles de decisión

La red clasifica correctamente al 70,56% de los registros y el árbol al 70,04%. Del árbol se puede inferir que la variable más importante para predecir el egreso es la nota de enseñanza media. La variables sede de la carrera no aparece como relevante. Para las reglas de asociación se utilizó la siguiente configuración (Figura 16):

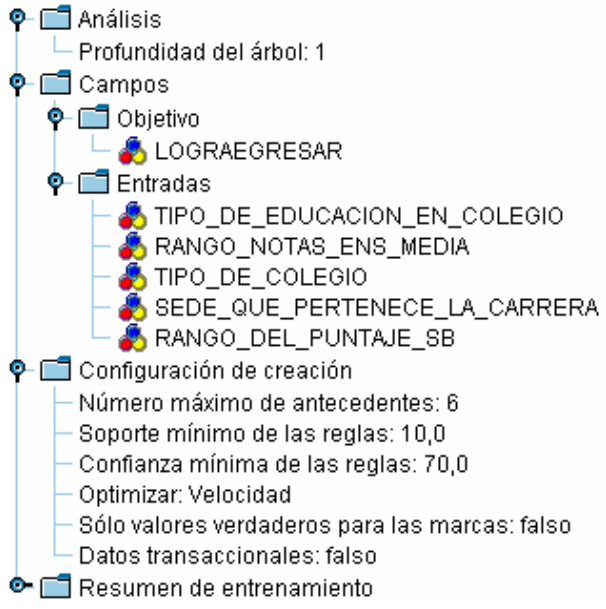

Figura 16. Resultados para caso Derecho con reglas de asociación

El resultado son 30 reglas para el atributo lograegresar= "no" entre los que se destacan los promedio de notas de enseñanza media inferiores a 5,99, el tipo de educación Humanista-Científico Nocturno y los tipos de colegios Particulares NO subvencionados y de la Corporación Municipal.

A pesar de que lo mayoría de los datos correspondían a alumnos de colegios particulares subvencionados, este tipo no aparece en ninguna regla.

Un análisis agregado de las cinco carreras analizadas se presenta en la Tabla 2.

Carrera	Condiciones de ingreso a evaluar	Mejor herramienta de DM	Razones
Arquitectura	*Nota de Enseñanza Media* *Tipo de Colegio, Tipo de Educación en colegio* *Puntaje de Selección*	Reglas de Asociación	Las reglas de asociación son una mejor herramienta debido a que permiten encontrar patrones en los datos, cosa que no se da con los árboles. Al inspeccionar las reglas entregadas para el caso de arquitectura es fácil darse cuenta de que no hay un patrón para la carrera de arquitectura, debido a que todos los valores posibles de los atributos aparecen en ellas.
Ingeniería Ejecución en Computación	*Nota de Enseñanza Media,* *Tipo de Colegio, Tipo de Educación en colegio* *Puntaje de Selección*	Reglas de Asociación	Al igual que arquitectura, hay un gran desbalance en los datos de egresados y eliminados, por lo que los árboles y las redes no son útiles. Algunas de los patrones que más se repiten para la eliminación, son los alumnos con nota promedio de enseñanza media inferior al 5,99; los puntajes de selección menores a los 600 puntos y los alumnos que hayan tenido un tipo de enseñanza en su colegio Humanista-Científica Diurna o Nocturna y Técnica Profesional Comercial.

Tabla 2. Resultados agregados del análisis usando Data Mining en las carreras en estudio.

Derecho	*Nota de Enseñanza Media,* *Tipo de Colegio, Tipo de Educación en colegio* *Puntaje de Selección*	Árboles de Decisión y Reglas de Asociación	En este caso los árboles nos dan una visión global de cómo se comporta la carrera, mientras que las reglas permiten encontrar algunos patrones de interés en los datos. El resultado son 30 reglas para el atributo *lograegresar*= "NO" entre los que se destacan los promedio de notas de enseñanza media inferiores a 5,99, el tipo de educación Humanista-Científico Nocturno y los tipos de colegios Particulares NO subvencionados y de la Corporación Municipal.
Periodismo	*Nota de Enseñanza Media,* *Tipo de Colegio, Tipo de Educación en colegio* *Puntaje de Selección*	Reglas de Asociación	Debido a la gran cantidad de egresados versus los eliminados, los árboles de decisión no pueden modelar esta carrera, por eso la mejor opción son las reglas. Este resultado muestra que las variable y los valores que deben alcanzar estas variables para predecir favorablemente el egreso son; promedio de notas de enseñanza media sobre 5,5; alumnos de colegios particulares subvencionados y NO subvencionados; alumnos de colegios de la corporación municipal y con tipo de enseñanza humanista científico diurno; alumnos con puntajes de selección sobre 500 puntos
Geología	*Nota de Enseñanza Media,* *Tipo de Colegio, Tipo de Educación en colegio* *Puntaje de Selección*	Reglas de Asociación	Debido a la gran cantidad de eliminados versus los egresados, los árboles de decisión no pueden modelar esta carrera, por eso la mejor opción son las reglas. El resultado da un total de 43 reglas para el atributo *estado del alumno*= "eliminado". Algunos de los patrones más repetidos para los atributos son el tipo de enseñanza Humanista-Científico Nocturno; el promedio de notas de enseñanza media inferiores a 5,99; el puntaje de selección inferiores 649,99 puntos y los colegios Particulares No Subvencionado y de la Corporación Municipal

Tabla 2. Resultados agregados del análisis usando Data Mining en las cinco carreras en estudio (Continuación).

CONCLUSIONES DEL ESTUDIO

La implementación de herramientas de data mining para apoyar el proceso de toma de decisiones y la formulación de políticas de admisión y retención de alumnos en la Universidad Católica del Norte permitió visualizar a través de resultados concretos la influencia que tienen ciertas variables de entrada sobre la permanencia de los alumnos en la universidad. Asimismo el uso de data mining facilitó el entendimiento de las variables que influencian la retención o pérdida de la calidad

de alumno o egresado de la universidad. La hipótesis planteada en este proyecto fue positivamente validada. Es decir que en general las condiciones iniciales de notas de enseñanza media, prueba nacional de ingreso a la universidad (PSU) y colegio de origen permiten predecir con diferentes matices si un estudiante finalizará exitosamente o no su carrera en la UCN. Sin embrago, la hipótesis no se pudo evaluar para el caso de la carrera de Arquitectura en particular ya que no presenta un patrón claro con las variables ingresadas en el estudio.

La hipótesis establecía que las condiciones de formación que tuvo el alumno en su enseñanza media permitían predecir si un alumno finalizaría exitosamente o no sus estudios de pre-grado. En este caso en particular la herramienta que entrego un resultado mas claro fueron los árboles de decisión ya que entrego con claridad que la variable carrera escogida y las notas de enseñanza media fueron las variables que explicaron en mayor mediada el egreso de los alumnos. Las redes neuronales se utilizaron inicialmente para estimar el modelo, pero estas no entregaron información de las variables de entrada que afectaban mayoritariamente el egreso. Los árboles de decisión en este caso también mostraron que los alumnos con nota enseñanza media menor que 6,0 no lograrían egresar.

Para el caso particular de las carreras en la UCN que no pueden ser modeladas mediante árboles de decisión, se vio que las reglas de asociación son una buena técnica para encontrar algunos patrones, como en el caso de Ingeniaría Ejecución en Computación e Informática y Geología, en la que se encontraron los patrones que tienen los alumnos eliminados de la carrera. Para el caso de Periodismo, fue posible encontrar, gracias a las reglas, las características de los alumnos que logran egresar de la universidad. Derecho es una carrera que fue satisfactoriamente modelada por un árbol de decisión, y además, se pudieron extraer algunos patrones adicionales mediante las reglas de asociación. Si bien para la carrera de arquitectura no se logró encontrar un patrón significativo y no se pudo modelar por los árboles, la conclusión que se obtiene es que las características de ingresos entregadas no predicen el éxito en esta carrera.

El desarrollo de este estudio utilizando la metodología CRISP DM entrego resultados positivos para la UCN ya que permitió sistematizar datos que existían en diversas fuentes y formatos. Las herramientas de data mining pueden presentar mayores ventajas que el uso de herramientas puramente estadísticas ya que ellas son de carácter básicamente exploratorio por lo que permiten trabajar con una mayor dimensionalidad del problema. En ese caso, se utilizaron árboles de decisión los que permitieron encontrar relaciones inéditas. Asimismo, las técnicas de data mining son menos restrictivas que las estadísticas, dado que no requieren por ejemplo condiciones de normalidad de datos y son tolerantes a ruidos en los datos.

Los resultados del estudio en el caso de la Universidad Católica del Norte utilizando herramientas de data mining sirvieron para focalizar los esfuerzos de retención de alumnos potencialmente más expuestos y proclives a la deserción, para mejorar la gestión de captura de alumnos con el mejor perfil posible para cada carrera de la UCN, para proponer políticas, procedimiento y acciones remediales que disminuyeran la deserción estudiantil, y para focalizar los esfuerzos en la promoción de la universidad teniendo más en claro a que colegios (y a que tipo de alumnos) atraer.

Finalmente, los resultados de este estudio permitirán a la UCN utilizar más eficientemente los recursos, detectando tempranamente causas que podrían implicar deserciones y que tienen un alto costo económico y social para el alumno, la institución, y el país.

LIMITACIONES DEL ESTUDIO

Este proyecto fue realizado en una Institución de educación superior de Chile cuyas características de financiamiento son particulares hacia las instituciones tradicionales autónomas con financiamiento del estado por lo tanto sus resultados son aplicables a instituciones de esta naturaleza. Por otra parte para su aplicación en otra institución de educación superior es necesario realizar la etapa de entendimiento de datos desde el inicio y considerando las condiciones particulares de la entidad en la que se aplicará. Es necesario destacar que la Universidad Católica del Norte posee sistemas de Información de tipo transaccional y data warehouse. Esta situación no es típica de las universidades chilenas por lo que este será un limitante para la aplicación de este proyecto en todas las universidades chilenas.

REFERENCIAS

1. Brachman, J.R. and Anand, T. (1996) The process of knowledge discovery in databases: a human-centered approach, In Fayyad, U., Piatesky-Shapiro, G., and Smyth, P *.Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, 37-58.

2. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000) CRISP-DM 1.0 Step-by-step data mining guide, SPSS Inc, USA. [Citado Febrero 2004]. http://www.crisp-dm.org/

3. Fayyad, U., Piatesky-Shapiro, G., and Smyth, P. (1996) Advances in Knowledge Discovery and Data Mining, The MIT Press, Cambridge.

4. Gartner Group (2004) Gartner Group, [Citado Febrero 2004]: http:// www.gartner.com

5. Gutierrez, M. (2004) Reporte Técnico Proyecto de Minería de Datos: Caso de Estudio en la Dirección General de Docencia de la Universidad Católica del Norte, Antofagasta, Chile.

6. Hand, D., Mannila, H., and Smyth, P. (2001) Principles of Data Mining, The MIT Press, Cambridge, MA.

7. Hernández, J., Mineya, N., and Montserrat, C. (2000) Extracción y Visualización de Conocimiento de bases de Datos Medicas, [Citado Febrero 2004]: http://www.acta.es/.

8. Jackson, J. (2002) Data Mining: A Conceptual Overview, *Communications of the Association for Information Systems*, 8, 267-296.

9. Luan, J. (2002) Data Mining, Knowledge Management in Higher Education, Potential Applications. Workshop and Presentation at *42nd Associate of Institutional Research International Conference*. Toronto , Canada.

10. Molina, L. (2002) Data Mining: Torturando los datos hasta que confiesen, Fundación Oberta de Cataluña (FUOC), Barcelona, España. [Citado Febrero 2004]: http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html.

11. Mujica, C. (2004) Necesidades de Información desde la Gestión Institucional. Visión desde una Vicerrectoria Académica, Reporte Técnico, Universidad Católica del Norte, Antofagasta, Chile.