Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación.

J. G. A. PAUTSCH, D. L. LA RED MARTINEZ, L. A. CUTRO

Dpto. Informática. FCEQyN. Univ. Nac. de Misiones
(3300) Posadas. Argentina
E-mail: gpautsch@fceqyn.unam.edu.ar

24 de Febrero de 2010

ABSTRACT

En la presente investigación se realizó una Minería de Datos sobre el Cubo 04 Desgranamiento, exportado del Sistema de Gestión Académica SIU-Guaraní, provistos por el Ministerio de Educación, Ciencia y Tecnología de la Nación. El objetivo principal fue maximizar la calidad que los modelos tienen para clasificar y agrupar a los estudiantes, de acuerdo a sus características académicas, factores sociales y demográficos, que han desertado de la Carrera Analista en Sistemas de Computación de la Facultad de Ciencias Exactas, Químicas y Naturales de la Universidad Nacional de Misiones. Luego, estos modelos fueron utilizados para realizar pronósticos sobre el resto de los alumnos. El proyecto se desarrolló bajo la metodología de libre difusión Crisp-DM y con la herramienta comercial IBM DB2 Warehouse (versión 9.5). La calidad de los modelos obtenidos a través de la clasificación con árboles de decisión superó a la técnica de agrupamiento a través de la generación de clústeres y ambas han superado ampliamente lo planteado.

Palabras Claves: Minería de datos, clasificación, agrupamiento, almacenes de datos, descubrimiento de conocimiento, deserción universitaria, perfiles de alumnos.

1 INTRODUCCIÓN

Todos los días, y casi sin darnos cuenta, se generan gran cantidad de datos informatizados. W.J. Frawley y otros[1], estiman que las bases de datos (BD) de las organizaciones se duplican cada veinte (20) meses. Por el contrario las técnicas de análisis de esta información no han tenido un desarrollo equivalente.

Muchas organizaciones mantienen grandes BD. Dentro de esta masa de datos hay información oculta de gran importancia que, aplicando procesos de Minería de Datos (MD) (data mining), se podría llegar a descubrir. (Figura 1.1)

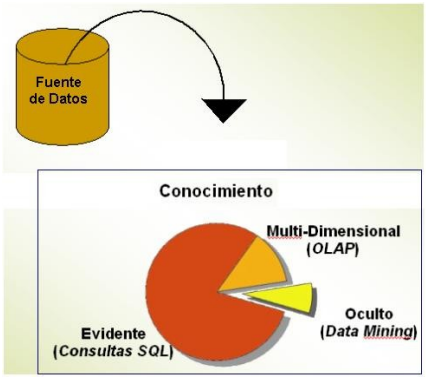


Figura 1.1. Tipos de conocimiento.

Se estima que un 80% de la información contenida en una BD corresponde al conocimiento evidente (fácilmente recuperable). El otro 20% requiere de técnicas más complejas para su obtención (Figura 1.2)

Puede que esta cifra parezca despreciable, pero la información oculta en ese pequeño porcentaje puede ser de vital importancia para el éxito de la empresa u organización.

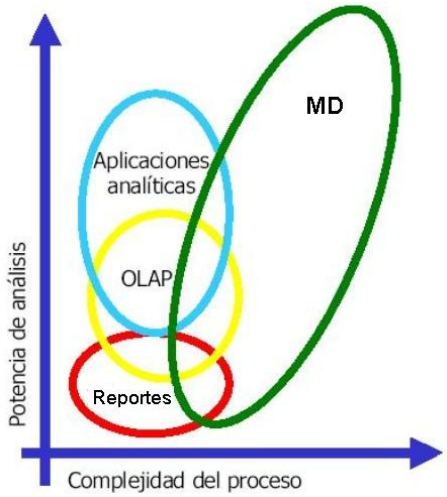


Figura 1.2. Relación entre potencial de análisis y complejidad.

Lo que se busca con esta tecnología es revelar conocimiento oculto útil y no evidente a partir de grandes BD [2].

Desde la década pasada la MD se ha ido incorporando a las organizaciones para constituirse en un apoyo esencial en el proceso de toma de decisiones.

2 OBJETIVO PRINCIPAL

Hoy día la Universidad Nacional de Misiones cuenta con el Sistema de Gestión Académica SIU-Guaraní (SIU-G). Su sigla significa **S**istema de **I**nformación **U**niversitaria y "es un consorcio de universidades que desarrolla soluciones informáticas y brinda servicios para el Sistema Universitario Nacional y distintos organismos de gobierno. Su objetivo es contribuir a mejorar la gestión de las instituciones, permitiéndoles contar con información segura, íntegra y disponible, optimizar sus recursos y lograr que el software sea aprovechado en toda su potencialidad." [18].

El sistema, produce una gran cantidad de datos, los cuales pueden ser muy valiosos, pero que resultan muy difíciles de analizar (debido a su volumen) por las autoridades; aún con el uso de las herramientas estadísticas, esta tarea se dificulta. Dentro de esta masa de datos hay información oculta de gran importancia que se podría llegar a descubrir con técnicas de MD.

Realizando un relevamiento preliminar, se observó que en el SIU-G existe un módulo (Interfaz) que exporta varios *Data Mart* de una DW. Los mismos están orientados al OLAP y abarcan diferentes temáticas.

Luego de analizar detalladamente la documentación que describe cada *Data Mart*[20], se determinó que el Cubo 04 – Desgranamiento, pueden ser de gran utilidad para la presente investigación, ya que aborda la temática de la deserción desde el punto de vista académico, social y demográfico.

El objetivo es realizar una MD, sobre las cohortes que se encuentran entre los años 2000 y 2006, a través de técnicas supervisadas y no supervisadas, sobre el Cubo 04 exportado de la BD del SIU-G. De esta forma se busca determinar cuáles son las técnicas, algoritmos y parámetros óptimos para extraer el conocimiento de la BD y así, confeccionar modelos para 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 de minería cuya calidad de predicción o clasificación supere el 65%. Por otra parte se buscará estandarizar y automatizar los procesos E.T.L. para que cada unidad académica pueda realizar la MD sobre el *Data Mart* exportado del SIU-G.

3 REVISIÓN CONCEPTUAL

La MD busca determina modelos compactos y comprensibles que rinden cuenta de las relaciones establecidas entre la descripción de una situación y un resultado.

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 [11] (Figura 3.1).

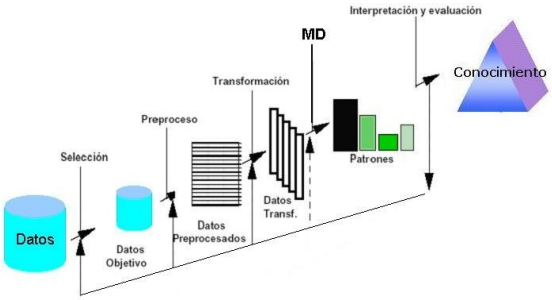


Figura 3.1. Etapas en el KDD

En la Figura 3.2 se pueden observar, algunas de las disciplinas que intervienen en la MD.

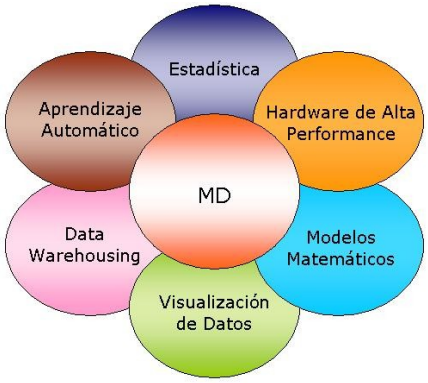


Figura 3.2. Disciplinas que intervienen en la Minería de Datos

El concepto de Data Mining no es nuevo. Desde los años 60, los estadísticos, manejaban términos como *Data Fishing*, *Data Mining* o *Data Archaeology*. La idea principal era encontrar correlaciones sin una hipótesis previa en BD con ruido.

Tampoco ninguno de los modelos estadísticos presentes en la MD son nuevos. Los árboles de decisión y de regresión (*classification and regression trees - CART*) son utilizados desde los años 60. Las bases de reglas fueron popularizadas durante el auge de los Sistemas Expertos en los 80 y las redes neuronales se conocen desde los años 40, pero han sido necesarios varios años de desarrollo para que fueran utilizables de manera sencilla.

Fue a principios de la década del 80 que Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de DM y KDD.

Ellos definen formalmente a la MD 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"[9].

La MD genera modelos que pueden ser descriptivos o predictivos[12].

- 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 SOFTWARE UTILIZADO

El software empleado para diseñar el modelo, crear la Almacén de Datos y realizar la MD fue IBM Data Warehouse Edition (DWE)V.9.5, que incluye al DB2 Enterprise Server Edition (DB2 ESE), al Design Studio (DS) y al Intelligent Miner (IM), cuyo permiso se encuentra autorizado por medio de las resoluciones N° 1417/04 y N° 858/06 de la Facultad de Ciencias Exactas, Naturales y Agrimensura (FACENA) de la Universidad Nacional del Nordeste y el "Acuerdo de Cooperación Tecnológica FACENA – IBM Argentina".

5 METODOLOGÍA

La metodología seleccionada fue CRISP-DM, ya que esta abarca una perspectiva más amplia contemplando también los objetivos empresariales del proyecto. Como reflejo de lo antes mencionado, podemos citar las primeras etapas de otras metodologías. Mientras que en ellas se comienza realizando un muestreo de datos, la metodología CRISP-DM se inicia realizando un análisis del problema de la empresa u organización, para su posterior transformación en un problema técnico[15].

La metodología CRISP-DM se acerca más al concepto real de proyecto, esto permite que pueda ser integrada con Metodologías de Gestión de Proyectos y así, completar las tareas administrativas y técnicas[16].

Otra diferencia significativa entre las metodologías radica en su relación con herramientas comerciales. La metodología SEMMA, por ejemplo, está ligada a los productos 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 (Figura 5.1), 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.

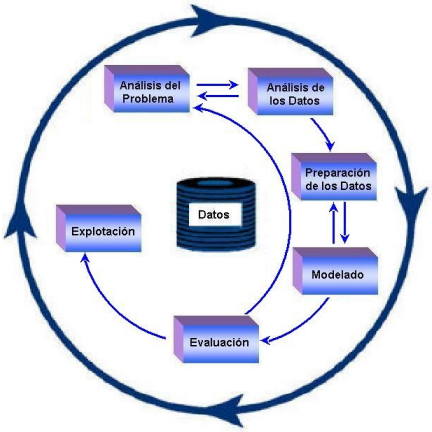


Figura 5.1. Fases del proceso de modelado metodología CRISP-DM.

6 RESULTADOS OBTENIDOS

En la Figura 6.1 se puede observar la estructura del Data Mart, Cubo 04 Desgranamiento, exportada del SIU-G.

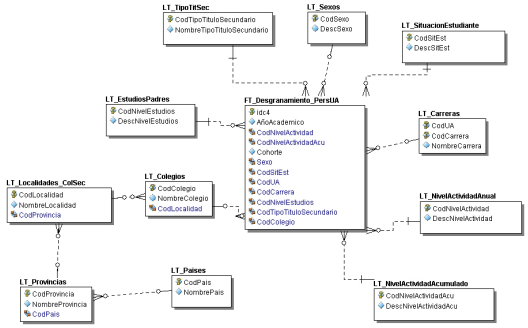


Figura 6.1. Cubo 04 - Desgranamiento

Se ejecutó el Flujo de Minería con la mejor configuración que se obtuvo en la etapa de Evaluación del Modelo, esto es:

- Nro. Clústeres: 14.
- Umbral Similitud: 90%.

Los resultados obtenidos se pueden observar en la Figura 6.2.

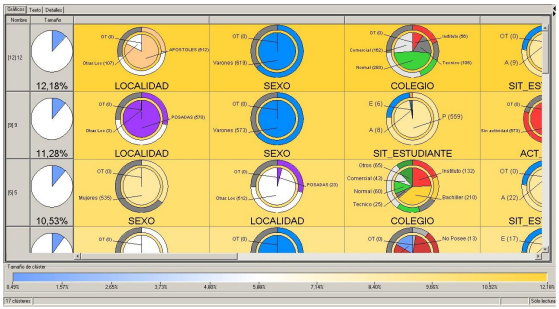


Figura 6.2. Gráfico de la Generación de Clústeres

En la tabla 6.3 se describe, en general, cómo están compuestos los primeros cinco Clústeres, los cuales agrupan más del 50% de la población.

Cluster	Atributo	Valor Predominante
Nro. 12 12,18% de la población	Localidad	Apóstoles
	Sexo	Varones
	Colegio	Normal
	Sit_Estudiante	P
	Act_Acu	Sin actividad
	Act_Anual	Sin actividad
	Provincia	Misiones
	Cohorte	2002
	Estudio_Padres	Pri
	Año_Acad	2008

Nro. 9 11,28% de la población	Localidad	Posadas
	Sexo	Varones
	Colegio	Instituto
	Sit_Estudiante	P
	Act_Acu	Sin actividad
	Act_Anual	Sin actividad
	Provincia	Misiones
	Cohorte	2000
	Estudio_Padres	Pri
	Año_Acad	2008
Nro. 5 10,53% de la población	Localidad	Otras Loc.
	Sexo	Mujeres
	Colegio	Bachiller
	Sit_Estudiante	P
	Act_Acu	Sin actividad
	Act_Anual	Sin actividad
	Provincia	Misiones
	Cohorte	2000
	Estudio_Padres	Pri
	Año_Acad	2008
Nro. 6 9,64% de la población	Localidad	Otras Loc.
	Sexo	Varones
	Colegio	Bachiller
	Sit_Estudiante	P
	Act_Acu	Sin actividad
	Act_Anual	Sin actividad
	Provincia	Misiones
	Cohorte	2000
	Estudio_Padres	Pri
	Año_Acad	2008
Nro. 11 7,60% de la población	Localidad	Apóstoles
	Sexo	Mujeres
	Colegio	Comercial
	Sit_Estudiante	P
	Act_Acu	Sin actividad
	Act_Anual	Sin actividad
	Provincia	Misiones
	Cohorte	2002
	Estudio_Padres	Pri
	Año_Acad	2008

Tabla 6.3. Descripción de los cinco Clústeres principales (50% de la población)

La Figura 6.4 muestra la calidad global del modelo. Esta es una medida de homogeneidad de los clusters. Su escala va de cero (0) a uno (1). De este modo un modelo cuya calidad global es uno (0) indica que las tuplas no tienen ninguna similitud con las demás tuplas de su cluster. Por el contrario un modelo cuya calidad global se aproxima a uno (1), indica que las tuplas del cluster son muy similares entre sí.

Una calidad global de 0,7 indica que, en promedio, las tuplas en un mismo cluster tienen en un 70% el mismo valor en los atributos activos.

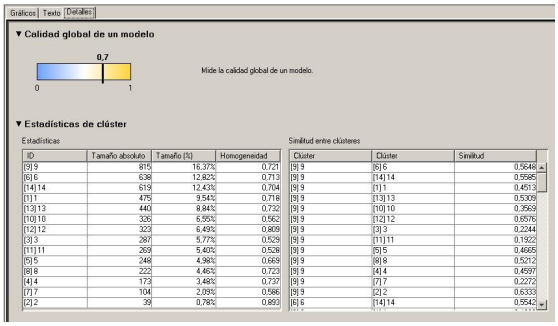


Figura 6.4. Calidad global del modelo obtenida con el algoritmo Generador de Clústeres.

Respecto a la clasificación con árboles de decisión, el Flujo de Minería se ejecutó con la mejor configuración que se obtuvo en la etapa de Evaluación del Modelo, que es la que ofrece la herramienta por defecto, esto es:

- Pureza máxima: 0.
- Profundidad máxima: 0.
- Número mínimo de registros por nodo hoja: 0.

Las clases que el algoritmo ha podido predecir se pueden observar en la Figura 6.5.

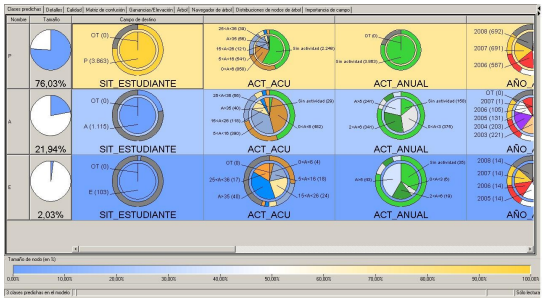


Figura 6.5. Clases predichas por el algoritmo Árbol

La tabla 6.6 describe, en general, cómo esta compuesta cada clase.

Clase	Atributo	Valor Predominante	Observaciones
"P" Alumnos Pasivos 76,03% de la población	Localidad	Otras Loc.	
	Sexo	Varones	Se mantiene la proporción con la población total
	Colegio	Comercial	Se mantiene la proporción con la población total
	Sit_Estudiante	P	
	Act_Acu	Sin actividad	
	Act_Anual	Sin actividad	
	Provincia	Misiones	
	Cohorte	2000	
	Estudio_Padres	Pri	Se mantiene la proporción con la población total
	Año_Acad	2008	
"A" Alumnos Activos 21,94% de la población	Localidad	Otras Loc.	
	Sexo	Varones	Se mantiene la proporción con la población total
	Colegio	Comercial e Instituto	Se mantiene la proporción con la población total
	Sit_Estudiante	A	
	Act_Acu	0<A<6 y 5<A<16	
	Act_Anual	0<A<3 y 2<A<6	
	Provincia	Misiones	
	Cohorte	2000	
	Estudio_Padres	Pri	Se mantiene la proporción con la población total
	Año_Acad	2008	
"E" Alumnos Egresados 2,03% de la población	Localidad	Posadas y Otras Loc.	Solo 7 registros de Apóstoles
	Sexo	Varones	Solo 7 registros de mujeres
	Colegio	Comercial, Instituto y Bachiller	
	Sit_Estudiante	E	
	Act_Acu	15<A<16 y A>35	
	Act_Anual	Sin actividad y A>5	
	Provincia	Misiones	
	Cohorte	2000	
	Estudio_Padres	Uni	
	Año_Acad	2008	

Tabla 6.6. Descripción de clases predichas por el algoritmo Árbol

La calidad global del modelo para clasificar a los alumnos Pasivos se puede observar en la Figura 6.7.

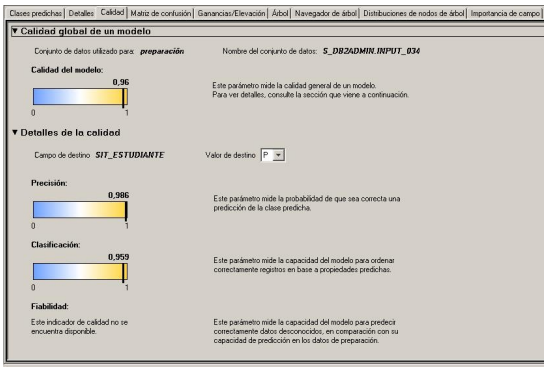


Figura 6.7. Calidad del modelo obtenida con el algoritmo Árbol

Las reglas establecidas por el árbol de decisión resultante se pueden apreciar en la Figura 6.8.

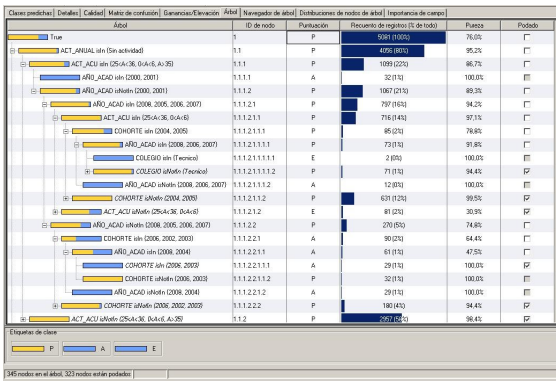


Figura 6.8. Árbol de Decisión.

Si bien en la ejecución de cada Flujo de Minería se mide la Calidad Global del modelo, para realizar la validación del modelo con datos reales, luego de seleccionar la Fuente de Datos, se introduce al área de diseño el elemento *División Aleatoria*. Este elemento permitirá dividir la Fuente de Datos en dos partes, una porción se utilizará para la construcción del modelo y la otra porción se dejará apartada para, una vez creado el modelo, validarlo. En la Figura 6.9 podemos observar en el elemento de División Aleatoria las dos salidas de datos:

- *Salida de Preparación*: es la que se utilizará para crear el modelo.
- *Salida de Prueba*: es la que se utilizará para validar el modelo.

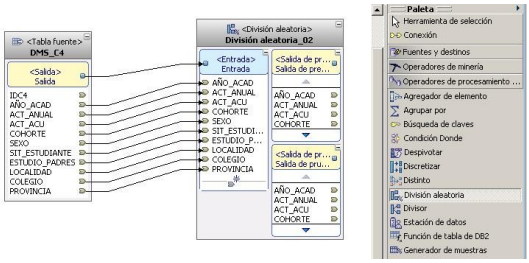


Figura 6.9. División Aleatoria de la Fuente de Datos

En referencia a la cantidad de datos definida para una u otra salida podemos decir que la proporción de error disminuye a medida que la cantidad de datos de entrenamiento aumenta. Para la presente evaluación se han realizado varias pruebas dando como resultado que la calidad del modelo más alta se logró con la siguiente configuración:

- *Salida de Preparación*: se utilizó el 67% de los datos.
- *Salida de Prueba*: se utilizó el 33% de los datos.

Paso seguido se introduce el elemento *Pronosticador*. Este es el encargado de crear el modelo de MD. La entrada de éste, debe conectarse a la Salida de *Preparación* del Divisor Aleatorio (Figura 6.10)

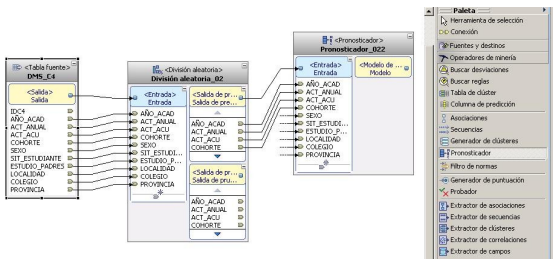


Figura 6.10. Conexión de la División Aleatoria con el Pronosticador

Para poder validar el modelo creado con los datos que se dejaron apartados en el paso anterior, debemos agregar al área de diseño un elemento *Probador*. Este artefacto tiene dos entradas:

- *Modelo de Minería*: que será sometido a prueba utilizando los datos de entrada. Debe conectarse al Modelo de Salida del elemento Pronosticador.
- *Entrada*: son los datos para realizar la prueba sobre el Modelo de Minería de entrada. Irá conectada a la Salida de Prueba del Divisor Aleatorio.

Por último, para visualizar el resultado de la prueba realizada, se puede introducir un elemento *Visualizer* y conectar su entrada al *Resultado de Prueba* del elemento Probador.

La validación para el Modelo de Clasificación, puede observarse en la Figura 6.11.

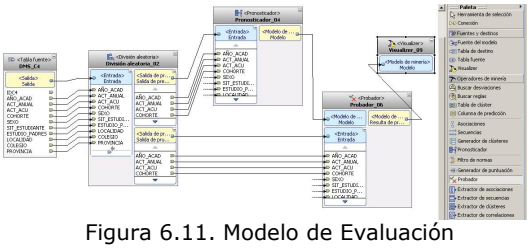


Figura 6.11. Modelo de Evaluación

Dentro del artefacto *Probador* es donde se deben hacer correlacionar los datos de entrada, provenientes de la muestra que se apartó para la probar el modelo, con los atributos del Modelo de Minería resultante del *Pronosticador* (Figura 6.12).

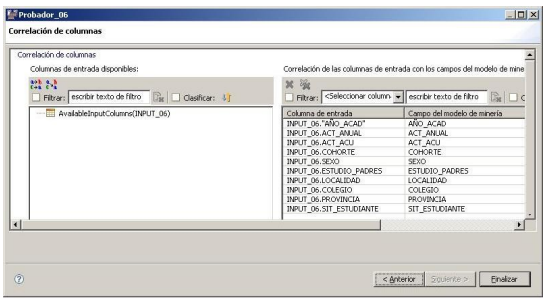


Figura 6.12. Parámetros Probador, correlación de Columnas

Si bien esto último, en este caso, puede parecer obvio, se debe tener en cuenta que la muestra para validar el Modelo de Minería puede provenir de otra fuente de datos, por ejemplo de alumnos que queremos pronosticar su futuro comportamiento, con encabezados diferentes. Aquí reside la potencia del modelo y de allí la importancia de trazar esta correlación entre los datos.

A continuación la Figura 6.13 muestra la calidad que el modelo tiene para clasificar los datos reales de los alumnos cuya *Sit_Estudiante='P'* (Alumnos Pasivos).

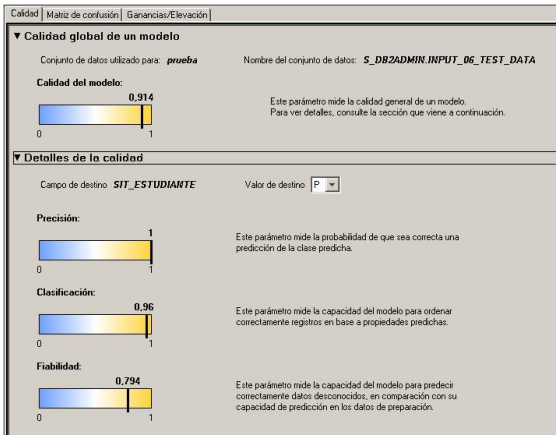
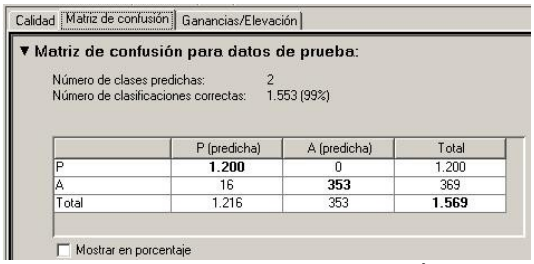


Figura 6.13. Calidad del Modelo

En la matriz de confusión de la Figura 6.14 podemos ver que el modelo clasificó incorrectamente sólo 16 tuplas de un total de 1.569.



	P (predicha)	A (predicha)	Total
P	**1.200**	0	1.200
A	16	**353**	369
Total	1.216	353	**1.569**

Figura 6.14. Matriz de Confusión

7 CONCLUSIONES Y LÍNEAS FUTURAS

La interpretación de los resultados se delego a los expertos en el dominio de la deserción. Todos ellos han observado que, si bien se realiza una buena clasificación y agrupamiento de las características de los alumnos Activos y Pasivos, salvo el Nivel de

Estudio de los Padres, la localidad, el desarraigo (*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.

No obstante en los resultados obtenidos con la generación de clústeres se pueden observar mayor nivel de alumnos Activos en los agrupamientos de estudiantes que vienen de localidades lejanas. Esto, por lo general, responde a que el estudiante, además de su convencimiento, deben convencer a su familia, obtener recursos, buscar dónde residir durante el período lectivo, enfrentar el desarraigo, etc. Esto marca otro nivel de compromiso comparado con los estudiantes que viven en la misma localidad donde se dicta la carrera, sumado a que, particularmente en Apóstoles, no existen muchas opciones para los estudiantes que quieren seguir una carrera universitaria.

Respecto al tratamiento de los colegios, se presenta el inconveniente de los colegios polimodales. Estos no poseen una orientación específica, ya que los contenidos los organizan cada institución en forma independiente. Así, por ejemplo, un colegio con orientación humanista, puede tener contenidos más asociados a ramas técnicas o matemáticas y, viceversa con los colegios de orientación técnica o comercial. Existen colegios en que, en los últimos dos años de cursado, no se han dictado matemáticas.

Un dato interesante que se puede apreciar en la generación de clústeres son los colegios comerciales de Apóstoles y Posadas que concentran la mayoría de los alumnos pasivos.

Una variable académica muy decisiva en la clasificación de los alumnos pasivos es la Actividad Anual. Según los modelos obtenidos, los alumnos que en un año no registran actividad académica (*Act_Anual = sin actividad*), luego no vuelven a registran ningún tipo de actividad, quedando pasivos. Este dato es muy interesante ya que acorta en un año la política de pasividad con la cual se manejan las unidades académicas hoy día.

Como respuesta al análisis de la temática abordada, los expertos concluyeron que, si bien se realiza una muy buena distinción entre los estudiantes activos y pasivos, no puede afirmarse que esta información sea concluyente para determinar si un alumno puede o no desertar de la carrera. Sería interesante que intervengan otras variables, sobre todo socio económicas en el estudio.

Como conclusiones del lado del ingeniero en conocimiento, primeramente se debe comentar que en esta investigación 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 (Flujos de Datos, Flujo de Minería, Matrices de Confusión, etc.) que faciliten y permiten realizar una validación de los modelos y un análisis de resultados más completo y fiable.

Con las dos técnicas seleccionadas 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).

El modelo de Clasificación a través de Árboles de Decisión superó en calidad a los patrones obtenidos con el método de Generación de Clústeres. A su vez, este ultimo, permitió 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, no resulta fácil de leer, debido a su amplitud, por personas no especializadas. Inclusive configurando distintos niveles de poda el árbol sigue siendo muy extenso. Esta dificultad es compensada, tanto en la clasificación como en el agrupamiento de características, por la excelente representación gráfica que realiza la herramienta.

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 Flujo de

Datos y Control. Con esta herramienta 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, y a la automatización de los flujos de datos, no representaría mayor inconveniente, el introducir más variables socio económicas, como sugieren los expertos.

A lo largo del desarrollo del presente proyecto han surgido varias líneas para ser abordadas en futuras investigaciones.

Entre algunas de ellas podemos mencionar:

- Investigar la manera en que el Flujo de Minería pueda ser incorporado al Flujo de Control para que, de esta manera, todo el proceso quede automatizado.
- Confeccionar los Flujos de Datos, Control y Minería de Datos para procesar los demás cubos provistos por la Secretaria de Políticas Universitarias (dependiente del Ministerio de Educación, Ciencia y Tecnología de la Nación).
- Diseñar nuevos cubos incorporando más variables socio económicas como estado civil, situación laboral, familiares a cargo y otras contenidas en la Base de Datos del SIU-G, particularmente en la tabla sga_Datos_Censales, y las sugeridas por los Expertos en la Sección 13.1. (por ejemplo, el desarraigo del núcleo familiar calculado en kilómetros).
- Desarrollar la presente investigación utilizando herramientas Open Source como por ejemplo la Suite Pentaho la cual provee e implementa todas las estructuras aquí vistas (Flujos de Datos, Control y Minería de Datos).

8 REFERENCIAS

[1] W.J. Frawley, G. Piatetski-Shapiro, C.J. Matheus, "*Knowledge Discovery in Databases*", AAAI-MIT Press, Menlo Park, California, 1991.

[2] Cabena P., Hadjinian P., Stadler R., Verhees J. & Zanasi, "*Discovering DataMining from Concept to Implementation"*, Book & Cd edition, September 1997.

[3] Silberschartz, Korth & Sudarshan, "*Fundamentos de Bases de Datos"*, Libro, Quinta Edicion, Mayo 2005.

[4] C.J. Date. "*Introducción a los Sistemas de Base de Datos"*, Libro, Séptima edición, 2001.

[5] Chaudhuri S. et al., "*An Overview of Data Warehousing and OLAP Technology*", Marzo 1997.

[6] Inmon W., "*Building Data Warehouse*", Technical Publishing Group 1992.

[7] Ramón García Martínez , Paola Verónica Britos, Alejandro Hossian, Enrique Sierra. *"Minería de datos Basada en Sistemas Inteligentes",* Primera edición, 2005.

[8] Kimball, R., "*The Data Warehouse Toolkit".* John Wiley & Sons, 1996.

[9] Frawley, Piatesky-Shapiro, Matheus, *"Knowledge Discovery in Databases: an Overview"*. AI Magazine, Otoño 1992.

[10] Luis Carlos Molina Félix, "*Torturando a los Datos Hasta que Confiesen*". Coordinador del programa de Data Mining, Universidad Oberta de Catalunya (UOC), 2001.

[11] Frawley, Piatesky-Shapiro, Smyth, "*From Data Mining to Knowledge Discovery in Databases*", 1996.

[12] Rakesh Agrawal & John C. Shafer: *"Parallel Mining of Association Rules"* IEEE Transactions on Knowledge and Data Engineering, December 1996.

[13] Sas Institute, Disponible en: http:// www.sas.com/technologies/analytics /datamining/miner/semma.html : Fecha de Consulta: Junio, 2009.

[14] Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler), *"CRISP-DM 1.0. Step-by-step data mining guide",* 1999.

[15] José E. Gondar, Consultoria de Datos, "Comparación de Metodologías de Data Mining", Disponible en: http://www.josebhuerta.com/datamining.htm, Fecha de Consulta: Julio, 2009.

[16] DataPrix, Disponible en: http://www.dataprix.com/el-modelo-crisp-dm-mineria-de-datos, Fecha de Consulta: Junio, 2009.

[17] Diario Digital Amsafe, entrevista realizada a Carlos Pallotti, presidente de la Cámara de Software y Servicios Informáticos de la República Argentina (Cessi), Diciembre, 2006.

[18] SIU Guarani, Disponible en: http://www.siu.edu.ar/acerca_de/que_es_el_siu
Fecha de Consulta: Septiembre 2009.

[19] Dean Abbott, "*An Evaluation of High-end Data Mining Tools for Fraud Detection*", IV IEEE International Conference on Systems, Man, and Cybernetics 1998.

[20] SIU Guarani, Descripción del Cubo 04 Desgranamiento, Secretaria de Políticas Universitarias dependiente del Ministerio de Educación, Ciencia y Tecnología de la Nación.

[21] Guía YPF, Disponible en: http://www.guiaypf.com.ar/guiaypf/ar_es/home/home.aspx
Fecha de Consulta: Octubre 2009.

[22] Lic. Mariana Inés Kubski, "*Minería de Datos con Intelligent Miner*", Universidad Nacional del Nordeste, Facultad de Ciencias Exactas, Naturales y Agrimensura, 2004.

[23] María N. Moreno García, Luis A. Miguel Quintales, Francisco J. García Peñalvo y M. José Polo Martín , "*Aplicación de Técnicas de Minería de Datos en la Construcción y Validación de Modelos Predictivos y Asociativos a Partir de Especificaciones de Requisitos de Software*", Universidad de Salamanca. Departamento de Informática y Automática, 2001.

[24] Servente, M. & García-Martínez, R., "*Algoritmos TDIDT Aplicados a la Minería de Datos Inteligente*". Facultad de Ingeniería. Universidad de Buenos Aires. 2. Director Adjunto del Programa de Magister en Ingeniería de Software. Escuela de Postgrado. Instituto Tecnológico de Buenos Aires, 2002.

[25] Mgter. Horacio Daniel Kuna, "*Memoria de Docencia e Investigación*", DEA Doctorado en Ingeniería de Sistemas y Computación, Universidad de Málaga, Agosto 2008.

[26] Mgter. David L. la Red Martínez, "Sistemas Operativos", sitio web: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/SOF.htm , Fecha de consulta 02/08/09.

[27] IBM Academic Initiative, Inicitaiva Académica de IBM para las Universidades del Mundo, sitio web: http://www-304.ibm.com/jct01005c/university/scholars/academicinitiative/, Fecha de consulta 12/10/09.