Aplicación de una nueva metodología Adaptive Business Intelligence para un análisis taxonómico predictivo utilizado para la detección temprana de alumnos universitarios en riesgo de deserción.

Roldán M. F. (1)

Application of a new methodology Business Adaptive Intelligence for predictive taxonomic analysis applied to the early detection of university students at risk of dropping out.

Abstract

In this paper I analyze the institutional challenges of digitalization and increasing use of the new information and communication technologies (ICTs) for the processes of knowledge management in Latin American public universities. The discussion of these challenges is organized in two parts; in part one I deal with the difficulties created by knowledge privatization and, in particular, because of the worldwide features of privatization. In the second part I analyze the institutional consequences of the international mobility of highly skilled workers. Finally, I conclude with an outline of a proposal that describes the strategies and institutional changes that would permit the Latin American public university to recover its visibility and reshape itself as a credible institutional actor by means of more pertinent and effective forms of knowledge management.

Key words: Adaptive Business Intelligence, Predictive Software, University Desertion, Methodology

Resumen

A partir de los datos de un caso de estudio de una Universidad local, se exponen los pasos para desarrollar una aplicación autoadaptativa de predicción, desde el modelo obtenido de una herramienta de extracción de conocimiento, utilizando las fases de una nueva metodología de Adaptive Business Intelligence, hasta la fase de Desarrollo del Software Predictivo. Para esto se han tomado en consideración los datos socio-económico-culturales de los alumnos ingresantes y su terminalidad de estudios. Con estos datos y aplicando una metodología de Adaptive Business Intelligence (ABI) recientemente creada y orientada a ciencias de la vida, se ha generado un modelo de aprendizaje que clasifica las causalidades de deserción o terminalidad en un contexto competitivo. La aplicación de la metodología en su fase final, culmina en una etapa denominada "Desarrollo de Software Predictivo", de la cual surge un modelo que puede ser aprovechado para generar aplicaciones orientadas a predicción, las que se podrán implementar en diferentes lenguajes de programación. En virtud de ello, se exponen las principales ventajas obtenidas de la explotación de las capacidades predictivas sobre los datos del caso y su aplicación con nuevos alumnos ingresantes a la universidad.

Palabras claves: Adaptive Business Intelligence, Software Predictivo, Deserción Universitaria, Metodología

(1) Dpto. Académico de Ciencias Exactas, Físicas y Naturales. Universidad Nacional de La Rioja, Luis M. de la Fuente, 5300 La Rioja, Argentina, e-mail: marcelo.roldan@unlar.edu.ar

Introducción

En los últimos años, ha existido un gran crecimiento de nuestras capacidades de generar y colectar datos, debido básicamente al gran poder de procesamiento de las máquinas con su bajo costo de almacenamiento. Dentro de estos enormes volúmenes de datos, existe una gran cantidad de información "oculta", de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información.

El presente estudio tiene como propósito exponer la aplicación de una nueva metodología de Adaptive Business Intelligence, la cual utiliza como datos de entrada un banco de datos proporcionado por una universidad nacional. Esta base de datos contiene la información respecto a los alumnos en diferentes aspectos, tales como datos personales, económicos, sociales, culturales, todos ellos favorecerán un análisis multivariado. Para este análisis se utilizaron diversas herramientas que permitieron el pre-procesamiento de los datos, su selección acorde a la sensibilidad con los resultados, clasificadores, discriminantes, normalizadores y técnicas de inteligencia artificial. Todo lo cual resulta en un árbol de clasificación que precisa aquellas causalidades principales de la deserción de los alumnos universitarios, en aquella institución mencionada.

Mediante una revisión sintética, se exponen además aquellos conceptos más relevantes para la aplicación posterior de las tecnologías en cuestión.

La investigación presentada se ha realizado sistemáticamente aplicando paso a paso la metodología ABI tal como lo indica su autor [11].

Entre los principales resultados se ha podido utilizar de manera clara y eficaz la metodología mencionada previamente, cuyos resultados han orientado gradualmente el tratamiento de la información oculta en las numerosas variables del caso. Este procesamiento exitoso, ha resultado en la determinación fehacientemente de aquellos parámetros cuyo impacto en la deserción es alto.

De manera indirecta, la resultante de este trabajo facilitará la construcción de un software que permita el ingreso directo de los datos, indicando de manera predictiva si se trata de un alumno potencialmente desertor.

De igual manera, la aplicación de las reglas obtenidas a la base de datos del SIU Guaraní, facilitará la detección de aquellos alumnos que actualmente se encuentran en riesgo de deserción, con la consecuente posibilidad de adopción de medidas correctivas en la institución.

El resto del trabajo se encuentra organizado de la siguiente manera. Las Secciones 1.1 a 1.4 presentan los fundamentos teóricos abarcando la tecnología de extracción de conocimiento, la metodología utilizada, la procedencia de los datos y las ventajas competitivas a lograr. En la Sección 2.1 aplicamos la metodología seleccionada al caso de estudio de predicción de deserción de los estudiantes. La sección 2.2 precisa detalles acerca de la calidad de la aplicación y en 2.3 detalla las características del modelo de clasificación obtenido. Finalmente, la Sección 3 cierra el trabajo con las conclusiones separando en 3.1 aquellas conclusiones relativas a la aplicación informática de las indicadas en 3.2 relacionadas con la deserción temprana de los alumnos.

1.1 Del Datamining al Adaptive Business Intelligence

El descubrimiento de esta información "oculta" es posible gracias a la minería de datos (data mining), que brinda un conjunto de técnicas sofisticadas para encontrar patrones y relaciones dentro de los datos.

Esto permite la creación de modelos, es decir, representaciones abstractas de la realidad, como parte del proceso de descubrimiento de conocimiento (KDDP, por su sigla en inglés) que se encarga, entre otras cosas, de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan significado a estos patrones encontrados.

Por otro lado, Business Intelligence [], [] provee beneficios que se aplican no solamente en los ámbitos empresariales, donde por demás está decir las ventajas que reditúa. Esta alternativa para la toma de decisiones se ve potenciada cuando se complementa con la característica autoadaptativa de las aplicaciones. En pocas palabras, Adaptive Business Intelligence (ABI) [] es la disciplina que combina la predicción, la optimización, y la capacidad de adaptación en un sistema capaz de responder a dos preguntas fundamentales: ¿Qué es probable que ocurra en el futuro? y ¿cuál es la mejor decisión en este momento?

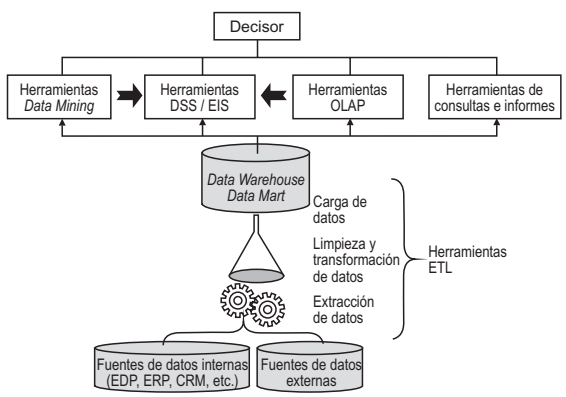

Figura 1. Arquitectura lógica de inteligencia empresarial (BI). Fuente [14]

En particular, los problemas de predicción se han convertido en un desafío para la extracción de conocimiento de la información, es así que una metodología basada en Adaptive Business Intelligence proporciona un conjunto de soluciones basadas en métodos y técnicas variadas (Data mining, Predicción, Optimización, y Adaptabilidad), las que permiten la extracción de conocimiento científico en la educación en particular así como en otras disciplinas en general.

Para implementar una tecnología como la que involucra al conjunto de técnicas descriptas previamente, se requiere de una metodología. [4]

1.2 La metodología de Adaptive Business Intelligence

Contar con una metodología, se ha convertido en algo tan importante y necesario como la carta de presentación de las empresas. Con esta premisa, se ha aplicado una nueva metodología para desarrollar un sistema de Adaptive Business Intelligence, a partir del estudio de numerosas variables que tienen correlación en mayor o menor medida con el indicador objetivo estudiado (deserción estudiantil).

La generación de la nueva metodología mencionada previamente a este trabajo, surge a partir de los conceptos relacionados entre disciplinas afines tecnológicamente como: OLAP, Acceso a datos multiplataformas, Business Intelligence, Data mining y Adaptabilidad.

Para su desarrollo, se han analizado las interrelaciones existentes entre ellas, y de este modo, se ha conformado un entorno de soporte para las aplicaciones predictivas.

Las etapas metodológicas para el desarrollo de aplicaciones basadas en Adaptive Business Intelligence abarcan la comprensión del problema, de los datos, de su preparación, modelado, búsqueda para acercarse a los objetivos e implementación a través de una aplicación de negocios.

A partir de estas etapas, la metodología propone el uso de las técnicas implementadas en la herramienta de minería de datos utilizada, buscando aquellos resultados que proporcionen la información necesaria para acercarse a los objetivos del proyecto. Esto involucra las etapas de "Búsqueda de patrones, reglas o grupos", la etapa de "Modelizado predictivo" y la "Validación del modelo".

Es en este punto donde esta metodología simple, ágil y efectiva, debería realizar los aportes necesarios para que la atención del investigador se vea favorecida, permitiendo una interacción dinámica con los patrones que surgen de los datos, a través de las diferentes fases, los cuales tienen significancia como nuevo conocimiento. [5]

Se exponen a continuación algunos conceptos relevantes, relacionados con esta nueva metodología, los que permitirán una comprensión más amplia de su uso en el caso de estudio tratado.

1.2.1 Ciclo de vida de la metodología

Al igual que sucede con otras metodologías, la sucesión de fases no es necesariamente rígida. Cada fase es estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones particulares, pero en ningún momento se propone cómo realizarlas.

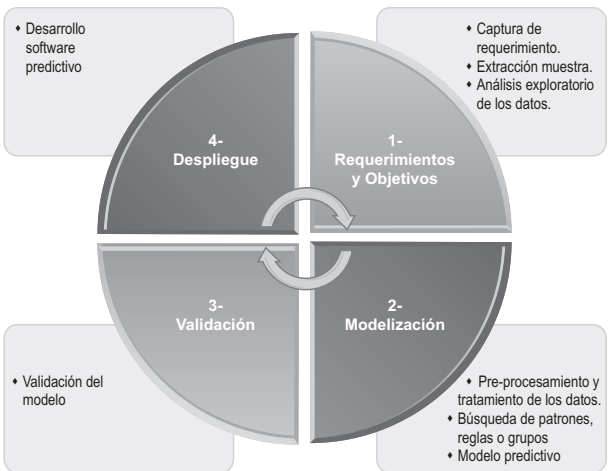

Figura 2 . Ciclo de vida de la nueva metodología Adaptive Business Intelligence y su correspondencia con las etapas metodológicas propuestas.

El ciclo de vida de la nueva metodología, mostrado en la Fig. 2, representa de manera sintética, aquellos aspectos que describen los pasos a seguir para lograr la realización de un modelo de despliegue, de tal manera que sea útil para un desarrollo de una aplicación predictiva.

Estos pasos son definidos de modo tal que permitan una evolución coherente y progresiva, que facilite, partiendo del problema planteado, llegar a una solución aceptable en los resultados esperados.

En esta misma figura se puede ver también la correspondencia con las etapas de la metodología. Dichas etapas, por su parte, permiten observar con un mayor grado de detalle un camino a seguir por el analista para encontrar los patrones ocultos en la información y en los datos, exponiendo las actividades que se desarrollan de manera secuencial durante el proceso metodológico de desarrollo de las aplicaciones autoadaptativas basadas en Business Intelligence.

Cabe notar que la fase 4 de Despliegue podría aparecer separada del resto, con la finalidad de representar el momento en el cual es posible la construcción del software [1]. Sin embargo, teniendo en cuenta que se plantea el carácter autoadaptativo para las aplicaciones que surgen de la utilización de esta metodología, este paso se vuelve cíclico junto al resto, ya que esta es la única manera de lograr que la aplicación optimice su predicción, alcanzando así el fin de Adaptive Business Intelligence, tal como lo indica Michaelewicsz: "Los sistemas de Adaptive Business Intelligence incluyen elementos de la minería de datos, modelos predictivos, predicción, optimización, y adaptabilidad, y son utilizados … para tomar mejores decisiones."[10]

En este contexto, el aporte del minero de datos como analista principal y de forma interdisciplinaria con el experto en el área de conocimiento, proveerá las nuevas fuentes que sustenten el mantenimiento futuro del sistema autoadaptativo en su conjunto. Para ello deberá reiniciar de manera rutinaria las actividades iterativas, a fin de detectar las nuevas tendencias en los datos o alteraciones significativas que pudieran variar la confiabilidad de los resultados predictivos. Cabe acotar que esto es posible de implementar en una aplicación predictiva, tal como lo hace Business Intelligence[11] a través de módulos específicos. Estos módulos pueden incorporar capacidades de análisis costo-impacto, análisis de matrices de costo y confusión, evaluación de hipótesis, técnicas de Boosting, Bagging, Randomization, y otras sofisticadas tecnologías.Sin embargo, la lógica estructural de la metodología contempla la implementación de la característica de autoadaptatividad y lo realiza semánticamente, tal como lo muestra la Fig. 3.

1.2.2 Pasos canónicos de la metodología

En esta sección presentamos de forma resumida el alcance de cada una de las etapas identificadas en la metodología ABI propuesta. Este nivel de detalle es necesario toda vez que la aplicación en el caso de estudio tal como se describe en los métodos, se resume a los resultados obtenidos en cada una de las etapas.

1. Captura de requerimientos. Como en cualquier proceso software, se trabaja estrechamente en el dominio de los expertos para definir el problema y determinar los objetivos del proyecto. Finalmente, los objetivos del proyecto se traducen en hipótesis acerca de la selección inicial de las técnicas de data mining que serán utilizadas más adelante en el proceso que se llevará a cabo.

2. Extracción muestral. Este paso incluye la recopilación de datos de la muestra y decidir qué datos, incluyendo el formato y tamaño, serán necesarios. Se comprueba la integridad de los datos, la redundancia, los valores que faltan, la plausibilidad de los valores de atributos, etc. Además, el paso incluye la verificación de la utilidad de los datos con respecto a los objetivos de data mining.

3. Análisis exploratorio de los datos. Ya que este tipo de sistemas son conducidos por datos, es importante tener una buena comprensión de estos. El objetivo es identificar los campos más importantes relacionados al problema y determinar cuáles valores derivados pueden ser útiles.Las etapas 2 y 3 pueden considerarse complementarias y pueden trabajar de manera conjunta en un ciclo correctivo de mejora.

4. Pre-procesamiento y tratamiento de los datos. En este paso se considera especialmente la limpieza de datos, que incluye la comprobación de la integridad de los registros de datos, la eliminación de o la corrección de ruido y los valores faltantes, remociones y duplicaciones, adecuaciones de formato, agregado de expresiones numéricas y fórmulas que mejoren o equilibren datos faltantes con mayor significancia o relevancia. En esta etapa, los datos podrán someterse a filtros supervisados o no supervisados que reduzcan los aspectos no afines a los objetivos del proyecto.

5. Búsqueda de patrones, reglas o grupos. En esta etapa se pretende que, a través de los diferentes métodos de data mining, los resultados obtenidos de la minería coincidan o se acerquen a los objetivos definidos en el paso 1. Es aquí donde se elige la técnica de minería de datos que permita encontrar patrones en los datos.

6. Modelado predictivo. Consiste en un refinamiento del método seleccionado en la etapa anterior, testeando diferentes técnicas para luego decidir cuál algoritmo y qué parámetros pueden ser utilizados de forma más minuciosa, de acuerdo a los requerimientos que el experto ha planteado.

7. Validación del modelo. Incluye la comprensión de los resultados, comprobando si el conocimiento descubierto es novedoso e interesante. Para ello la interpretación de los resultados por expertos en la materia es significativa cuando se trata de verificar el impacto del conocimiento descubierto. Debido a su carácter iterativo, es posible replicar las etapas anteriores utilizando otros métodos y algoritmos de data mining, identificando qué acciones alternativas se pueden adoptar para mejorar los resultados.

8. Desarrollo software predictivo. Una vez que se ha conseguido el modelo, para que este sea útil, deberá ser desplegado. Esto significa salir del entorno de desarrollo en una forma que pueda ser usada por un software externo. Esta etapa final consiste en la planificación de cómo se utilizará el conocimiento descubierto si fuera necesario.

Se presenta como una etapa final de esta metodología, pero como una etapa inicial de un proceso de desarrollo software, ya que plantea los requerimientos de un nuevo producto software, el cual utiliza las reglas de inferencia, clasificadores, árboles o informaciones que han surgido como conocimiento de la aplicación desarrollada a través de los pasos metodológicos descriptos hasta ahora.

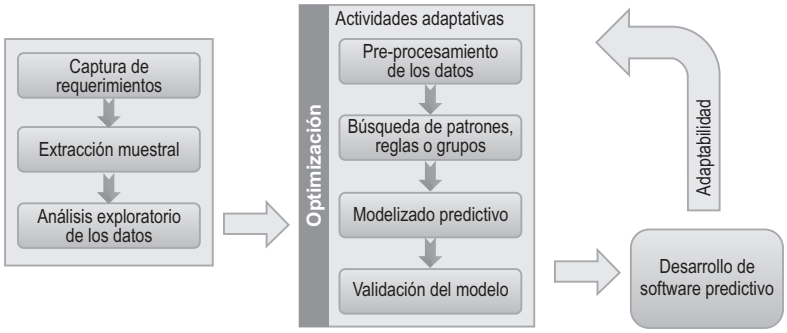

Figura 3 . Implementación de la optimización, predicción y adaptabilidad en la metodología.

1.3 Gestión de los datos de los alumnos – SIU Guaraní

Las universidades utilizan sistemas de información adecuados a sus necesidades y características propias tales como estructura de los planes, modalidad de cursado, constitución de sus sedes geográficamente entre otras características. Aunque existen diferencias notables entre las distintas instituciones, es posible mantener criterios comunes para el desarrollo de sistemas informáticos que permitan la gestión de los datos de los alumnos desde sus primeros días en la universidad hasta que culminan sus estudios. La iniciativa del SIU Guaraní es un ejemplo de esta tendencia a la unicidad de elementos que favorezcan los desarrollos informáticos. Estos sistemas de información tienen como uno de sus objetivos asegurar la protección y disponibilidad de la información, fiabilidad, sencillez y costo entre otras ventajas. [14]

El SIU Guaraní como sistema de información permite el seguimiento de todas las actividades que realiza un estudiante: inscripción a exámenes y cursados, reinscripción a carreras, consulta de inscripciones, consulta de plan de estudios e historia académica, consulta de cronograma de evaluaciones parciales, consulta de créditos, notas de evaluaciones parciales, materias regulares, actualización de datos censales y recepción de mensajes.[13]

Entre los usos parametrizables que las universidades pueden utilizar, el SIU-Guaraní se presenta como un sistema informático de autogestión acádémica por Internet, lo que permite a los alumnos un uso más adecuado de las tecnologías de la información actuales y a la institución una recopilación de mayor cantidad de datos.

Esta última ventaja es posible debido a que no se requiere uso de equipamientos dedicados en la institución, lo que limitaría el tiempo y el acceso de los alumnos a las computadoras. De esta manera es posible confeccionar un modelo de datos más abarcativo, el cual puede contener datos de mayor relevancia a diferentes contextos de análisis.

En la figura 4 y 5 se exponen dos pantallas donde se resumen datos de un alumno. Con los datos recopilados por el SIU Guaraní se construye la base de datos utilizada por una universidad argentina, aportando información de orden socio-económico-cultural de cada uno de los estudiantes. Esta base de datos constituye la referencia inicial para el procesamiento de los mismos mediante herramientas ETL (Extract-Transform-Load) que constituirán los datos de entrada para el análisis de patrones.

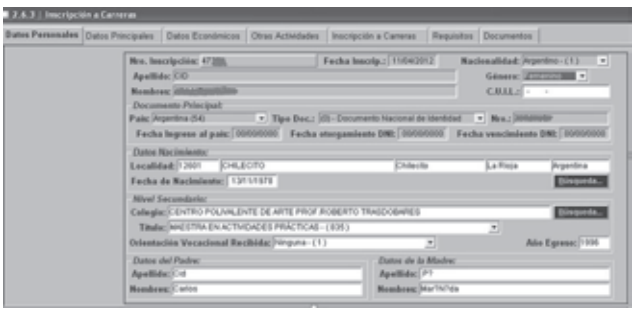

Figura 4. Ficha de datos personales del alumno. Interface con el usuario del SIU Guaraní.

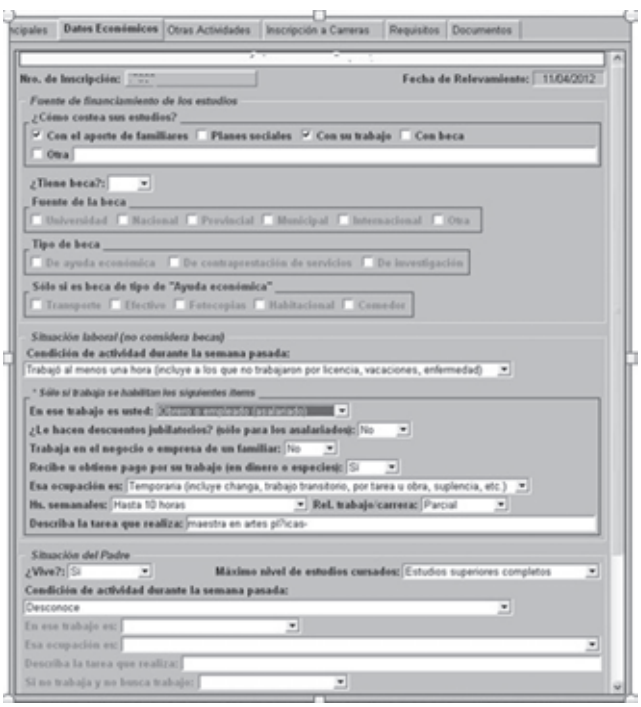

Figura 5 . Ficha de datos económicos del alumno. Interface con el usuario del SIU Guaraní.

1.4 Implementación estratégica de la aplicación

El uso de herramientas como la planteada en el presente trabajo frente a las amenazas externas, como lo son la aparición de nuevos actores en el escenario educativo, se vuelve relevante para las universidades públicas como estrategia competitiva.

La posibilidad de facilitar al decisor información precisa y confiable respecto de indicadores como el porcentaje de alumnos que han desertado, es una medida que se comporta como una medición pasiva, toda vez que cualquier acción correctiva adoptada dependerá de una política y a su vez tendrá un destino más amplio que preciso. Es allí donde el aporte de las tecnologías de Adaptive Business Intelligence puede proveer información oportuna y selectiva que permita un proceso de toma de decisiones óptimas mejorando la celeridad y la precisión. La implantación de herramientas predictivas se constituye de esta manera en una aplicación tecnológica cuyo impacto estratégico en la universidad pasa a ubicarla de un cuadrante de apoyo a una nueva situación de Impulsora de acuerdo a la Matriz de McFarlan.

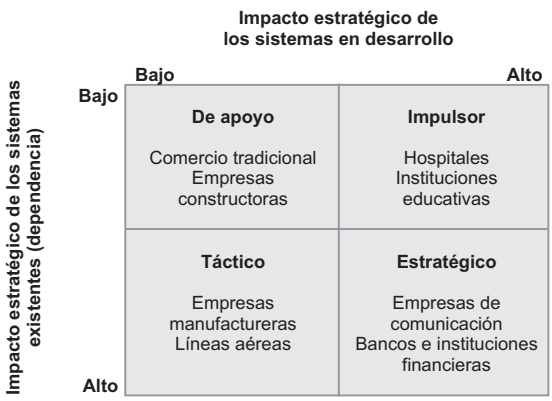

Figura 6. Matriz de McFarlan.

La matriz de McFarlan de la figura 6 permite analizar la situación actual de los sistemas de información y su proyección futura en relación con su importancia y relevancia estratégica para la institución. [14].

Materiales y métodos

2.1 Aplicación de la metodología para la predicción de casos de deserción potencial de los alumnos ingresantes. Los datos empleados provienen de una Universidad Nacional de Argentina. En ellos se encuentran alumnos en diferentes estadios de sus estudios con características sociales diversas. Se aborda su resolución utilizando Weka como herramienta [8].

Previo a la utilización de los datos aportados por la base de datos del SIU Guaraní, ha sido necesario realizar una selección empírica de aquellos atributos cuya incidencia puede ser relevante, dejando de lado algunos como: datos de tarjetas de crédito, identificadores de tipos de documentos, fechas irrelevantes, códigos, nombres, entre otros.

Fase 1: Captura de requerimientos: Resultados esperados al finalizar el proyecto

Se ha definido que el modelo realice la clasificación de los alumnos con riesgo de deserción con una precisión cuyo error absoluto medio sea <=2% y su precisión de al menos 80%.

Aspectos relevantes del proyecto

Características del conjunto de datos	Multivariable
Características de los atributos	Categóricos Numéricos
Tareas asociadas	Clasificación
Número de muestras inicial	4707
Número de atributos	47
¿Valores faltantes?	Si
Área de aplicación	Educación
Comparación de modelos	Si
¿Requiere desarrollo predictivo?	No

Tabla 1: Aspectos relevantes.

Atributo	Reemplazada por	Registros sin datos
nombre_alumno		0
nro_documento		0
fecha_inscripcion		0
nacionalidad		0
fecha_nacimiento	Edad=24	4
Edad	24	4
localidad_nacimiento		0
provincia_nacimiento		0
colegio_secundario	?	2715
titulo_secundario	?	2705
orientacion_recibida	"Ninguna"	2862
estado_civil	"Soltero"	99
vive_unido_de_hecho	N	30
cant_hijos	0	1227
obra_social		0
residencia_tipo	Otros	1123
con_quien_vive	"En otra situación"	1263
costea_estudios	?	1063
tiene_beca	N	801
situacion_laboral		0
padre_vive		0
max_est_cur_padre	?	330
madre_vive		0
max_est_cur_madre	?	167
DISP__PC_EN_CASA		0
DISP__PC_EN_TRABAJO		0
DISP__PC_EN_UNIVERSIDAD		0
DISP__PC_EN_OTRO_LUGAR		0
habla_ingles		0
nombre_carrera		0
sexo		0
carrera_cod		0
plan		0
legajo		0
Promedio		464
mat_aprob		0
fech_prim_exa		508
fech_ult_exa		508
longevidad_alumno		0
class	Abandono/Cursante	0

Tabla 2: Datos faltantes.

Se provee, al finalizar el proyecto, de aquellas reglas necesarias para la construcción de un software capaz de emular el comportamiento del modelo obtenido a partir de los datos de la base de datos.

Fase 2: Extracción muestral

Se trabajó con un banco de datos cuya información contiene un elevado número de muestras (4707 registros) a partir de los datos provistos por el SIU Guaraní de una Universidad Nacional de Argentina.

Clase	Número de muestras
Abandono	2456
Cursante	2251

Tabla 3 : Distribuciones por clases.

Esta base de datos contiene 45 atributos, 2 atributos adicionales han sido calculados en base a valores contenidos en otros atributos. Se determinó la condición de Cursantes o Abandonos a partir de aquellos alumnos que no han rendido asignaturas desde hace 2 años.

Los atributos restantes constan de valores numéricos en 6 atributos, y los restantes son nominales (discretos).

Fase 3: Análisis exploratorio de los datos

Información de los atributos. Distribución de clases

Base de datos: Facilitada por el Sistema unificado universitario de Gestión de Alumnos – SIU Guaraní de una Universidad Nacional de Argentina.

Ya que este tipo de sistemas son conducidos por datos, es importante tener una buena comprensión de los mismos (Tabla 4). El objetivo del modelo del data mining ha sido identificar los campos más importantes relacionados al problema y determinar cuáles valores derivados pueden ser útiles.

Hay ilimitadas maneras de visualizar datos, pero las dos herramientas fundamentales son el gráfico X-Y, el cual mapea relaciones entre variables, y el histograma, el cual muestra la distribución estadística de los datos. Ver figura 7.

En esta etapa, la exploración de la información ha simplificado el problema, con el objetivo de optimizar la eficiencia del modelo, siendo este el foco de atención de esta etapa de la metodología. Idealmente, se puede tomar todas las variables/características que se necesita y usarlas como entrada, para luego descartar las innecesarias.

Por lo tanto esta etapa se orienta mayormente hacia la visualización de los datos, con la finalidad de simplificar el problema, detectando aquellos datos con poca o ninguna incidencia estadística hacia los objetivos predefinidos.

fieldname	Type	length	precision	step origin	storage	mask
nombre_alumno	string	-		datos de alumnos		
nro_documento	number	-		datos de alumnos		#
fecha_inscripcion	date	-		datos de alumnos		yyyy/mm/dd
nacionalidad	string	-		datos de alumnos		
fecha_nacimiento	date	-		datos de alumnos		yyyy/mm/dd
Edad	integer	-	0	datos de alumnos		#
localidad_nacimiento	string	10		datos de alumnos		
provincia_nacimiento	string	10		datos de alumnos		
colegio_secundario	string	10		datos de alumnos		
titulo_secundario	string	10		datos de alumnos		
orientacion_recibida	string	-		datos de alumnos		
estado_civil	string	-		datos de alumnos		
vive_unido_de_hecho	string	-		datos de alumnos		#
cant_hijos	integer	-	0	datos de alumnos		
obra_social	string	16		datos de alumnos		
residencia_tipo	string	-		datos de alumnos		
con_quien_vive	string	10		datos de alumnos		
costea_estudios	string	20		datos de alumnos		
tiene_beca	string	-		datos de alumnos		
situacion_laboral	string	-		datos de alumnos		
padre_vive	string	-		datos de alumnos		
max_est_cur_padre	string	-		datos de alumnos		
madre_vive	string	-		datos de alumnos		
max_est_cur_madre	string	-		datos de alumnos		
DISP__PC_EN_CASA	string	-		datos de alumnos		
DISP__PC_EN_TRABAJO	string	-		datos de alumnos		
DISP__PC_EN_UNIVERSIDAD	string	-		datos de alumnos		
DISP__PC_EN_OTRO_LUGAR	string	-		datos de alumnos		
accede_internet_casa	string	-		datos de alumnos		
accede_internet_trabajo	string	-		datos de alumnos		
accede_internet_universidad	string	-		datos de alumnos		
accede_internet_cyber	string	-		datos de alumnos		
accede_internet_otro_lugar	string	-		datos de alumnos		
regularidad_accede_a-	string	-		datos de alumnos		
practica_deportes	string	-		datos de alumnos		
habla_ingles	string	-		datos de alumnos		
nombre_carrera	string	-		datos de alumnos		#
sexo	string	-		datos de alumnos		
carrera_cod	string	-		datos de alumnos		#,#
plan	string	-		datos de alumnos		yyyy/mm/dd
legajo	string	-		datos de alumnos		yyyy/mm/dd
Promedio	number	-		datos de alumnos		#
mat_aprob	integer	-	0	datos de alumnos		
fech_prim_exa	string	-		datos de alumnos		
fech_ult_exa	string	-		datos de alumnos		
longevidad_alumno	integer	-	0	datos de alumnos	normal	

Tabla 4: Detalle de los atributos.

Por lo tanto esta etapa se orienta mayormente hacia la visualización de los datos, con la finalidad de simplificar el problema, detectando aquellos datos con poca o ninguna incidencia estadística hacia los objetivos predefinidos.

Fase 4: Pre-procesamiento y tratamiento de los datos

Para la realización del experimento, se han revisado los datos encontrándose las relaciones mostradas en la figura 7 entre ellos.

Esta distribución de algunos datos, denota la posibilidad de realizar una selección de un subconjunto de datos, basados en algunas correlaciones posibles de estimar previamente a la modelización.

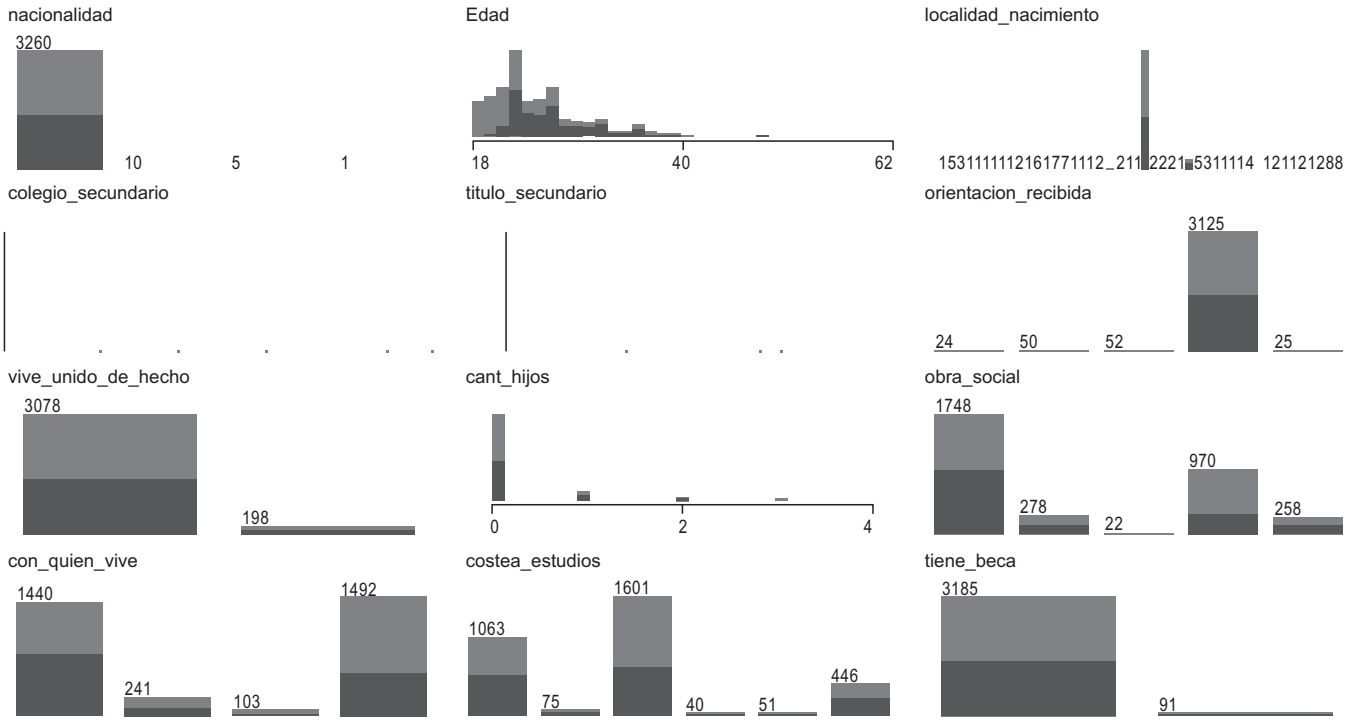

Figura 8 : Relaciones entre los datos procesados.

De esta manera, se han encontrado los siguientes atributos cuyas modificaciones son relevantes:

Edad: Existencia de registros con edad 0. Cambiado a la media de la edad.

Título secundario: 4117 valores faltantes. Se eliminó el atributo.

Orientacion_recibida: 4293 valores faltantes. Se eliminó el atributo.

Residencia_tipo: 2551 valores faltantes. Se reemplazó el faltante por 'Otros'

Con_quien_vive: 2498 valores faltantes. Se reemplazaron por 'En otra situación'

Costea_sus_estudios: 2262 valores faltantes. Se reemplazaron faltantes por 'N'

Tiene_beca: 801 valores faltantes. Se reemplazaron datos faltantes por 'N'

Padre_vive: Existe un grupo de alumnos (256) que han respondido D (Desconoce si su padre vive, se trate de hijos de madre soltera o de padre desconocido), en tales casos se asume que 'No' (como si la respuesta hubiera sido 'No vive'). Situación similar se presenta para Madre_vive (105), resolviéndose de manera idéntica.

Existe un grupo de 1.431 alumnos que no han respondido a la mayoría de los parámetros solicitados, entre algunos de los mencionados anteriormente, por lo que se los filtra para analizar el conjunto de datos restantes con mayor precisión.

Posterior a este filtrado de los datos, permanecen algunos atributos que requieren una modificación para evitar desvíos estadísticos.

Los siguientes datos corresponden a aquellos que se trataron luego mediante algún algoritmo de preprocesamiento:

Atributo	Reemplazada por	Registros sin datos
fecha_nacimiento	Edad=24	4
Edad	24	4
colegio_secundario	?	2715
titulo_secundario	?	2705
orientacion_recibida	"Ninguna"	2862
estado_civil	"Soltero"	99
vive_unido_de_hecho	N	30
cant_hijos	0	1227
residencia_tipo	Otros	1123
con_quien_vive	"En otra situación"	1263
costea_estudios	?	1063
tiene_beca	N	801
max_est_cur_padre	?	330
max_est_cur_madre	?	167
practica_deportes	N	746

Tabla 5 : Detalle de los atributos que requieren pre-procesamiento.

Atributo modificado	Valor	Nuevo valor	Acción realizada sobre el atributo
	original		
Longevidad_alumno	-	-	Marcado
Colegio_secundario	-	-	Marcado
Titulo_secundario	2705 = ?	-	Marcado
	-	-	Eliminación
Sexo	-	-	Marcado
	-	-	Discretización
			(unsupervised.atribute.discretize)

Tabla 6 : Acciones realizadas sobre los atributos

Respecto a los filtros de selección de atributos con mayor correlación a las clases finales de la clasificación, se han evaluado varios métodos cuyos resultados se exponen a continuación en la Tabla 7:

Subset	Algoritmo	Método de búsqueda	Atributos resultantes
1	-	-	Todos los atributos se procesan
2	Cfs Subset Eval	Best first	Selected attributes: 2,5,33,34 : 4 **Edad** **orientacion_recibida** **mat_aprob** **longevidad_alumno**
3	Cfs Subset Eval	Genetic search	Selected attributes: 2,5,29,32,33 : 5 **Edad** **orientacion_recibida** **practica_deportes** **Promedio** **mat_aprob**
4	InfoGain Attribute Eval	Attribute ranking	Ranked attributes: **0.293519 33 mat_aprob** **0.194954 2 Edad** **0.066067 32 Promedio** **0.041213 5 orientacion_recibida** **0.034228 10 residencia_tipo** **0.032234 9 obra_social** 0.022474 28 regularidad_accede_a_internet 0.020665 30 habla_ingles 0.020275 11 con_quien_vive 0.019837 3 localidad_nacimiento 0.016946 19 DISP__PC_EN_CASA 0.015368 23 accede_internet_casa 0.014423 26 accede_internet_cyber 0.010887 14 situacion_laboral 0.009985 27 accede_internet_otro_lugar 0.009691 4 provincia_nacimiento 0.008011 12 costea_estudios 0.006888 22 DISP__PC_EN_OTRO_LUGAR 0.006066 29 practica_deportes 0.005578 18 max_est_cur_madre 0.004556 25 accede_internet_universidad 0.004519 16 max_est_cur_padre 0.002976 1 nacionalidad 0.002443 8 cant_hijos 0.001761 21 DISP__PC_EN_UNIVERSIDAD 0.001204 17 madre_vive 0.000877 24 accede_internet_trabajo 0.000739 20 DISP__PC_EN_TRABAJO 0.000607 6 estado_civil 0.000245 13 tiene_beca 0.00018 15 padre_vive 0.000136 7 vive_unido_de_hecho 0.000121 31 sexo_1

Tabla 7 : Filtros de selección de atributos.

El análisis utilizando el algoritmo de filtrado InfoGainAttributeEval [8] a través de un método de ranking, permitió clasificar el ranking de atributos obteniendo las sensibilidades del Subset Nº 4.

Fase 5: Búsqueda de patrones, reglas o grupos

La metodología propuesta aprovecha en este caso las bondades de un algoritmo de clasificación en árbol SimpleCart, J48 y Decision Table, los que proveen un método supervisado para la clasificación.

Aplicados se obtienen los resultados de la tabla 8.

Paso	Algoritmo utilizado	Porcentaje de aciertos
1	**SimpleCart (Subset 1)**	Correctly Classified Instances 2833 86.47 % Incorrectly Classified Instances 443 13.52 %
2	**J48 (Subset 1)**	Correctly Classified Instances 2840 86.69 % Incorrectly Classified Instances 436 13.30 %
3	**Rules.DecisionTables (Subset 1)**	Correctly Classified Instances 2782 84.92 % Incorrectly Classified Instances 494 15.07 %
4	**SimpleCart (Subset 2)**	Correctly Classified Instances 2777 84.76 % Incorrectly Classified Instances 499 15.23 %
5	**J48 (Subset 2)**	Correctly Classified Instances 2791 85.19 % Incorrectly Classified Instances 485 14.80 %
6	**Rules.DecisionTables (Subset 2)**	Correctly Classified Instances 2774 84.67 % Incorrectly Classified Instances 502 15.32 %
7	**SimpleCart (Subset 3)**	Correctly Classified Instances 2856 87.18 % Incorrectly Classified Instances 420 12.82 %
8	**J48 (Subset 3)**	Correctly Classified Instances 2852 87.05 % Incorrectly Classified Instances 424 12.94 %
9	**Rules.DecisionTables (Subset 3)**	Correctly Classified Instances 2802 85.53 % Incorrectly Classified Instances 474 14.46 %
10	**SimpleCart (Subset 4)**	**Correctly Classified Instances 2858 87.24 %** **Incorrectly Classified Instances 418 12.75 %**
11	**J48 (Subset 4)**	Correctly Classified Instances 2854 **87.11** % Incorrectly Classified Instances 422 12.88 %
12	**Rules.DecisionTables (Subset 4)**	Correctly Classified Instances 2816 85.95 % Incorrectly Classified Instances 460 14.04 %

Tabla 8: Resultados obtenidos en la clasificación de los subconjuntos de datos.

Con esta aplicación de la clasificación, queda demostrado el correcto uso de la técnica de selección de datos basados en correlatividad. Es decir, los porcentuales de error y las precisiones alcanzadas, son mejores con los atributos recomendados por el evaluador de atributos con el método de ranking [8].

2.2 Aspectos de calidad del modelo

Fase 6: Modelado predictivo

Para el proceso de entrenamiento del modelo se ha utilizado la técnica de análisis denominada cross-validation (validación cruzada); habiendo aplicado los filtros “CFS Subset Evaluator” con los métodos “Best first” y “Genetic search”, además del filtro “Info Gain Attribute” con el filtro “Eval Attribute ranking” (todas ellas herramientas incorporadas en weka)[8], para eliminar los atributos con menor significancia. Luego de ello, el porcentaje final de acierto asumido como el mejor se muestra en tabla 9.

```
=== Summary ===

Correctly Classified Instances        2858       87.2405 %
Incorrectly Classified Instances      418        12.7595 %
Kappa statistic                          0.7434
Mean absolute error                      0.2025
Root mean squared error                  0.324
Relative absolute error                 40.8993 %
Root relative squared error             65.1173 %
Total Number of Instances             3276
```

Tabla 9 : Resultado del modelado predictivo.

Fase 7: Validación comparativa del modelo

Las figuras 9 y 10 muestran las evaluaciones del modelo realizadas a través del área bajo la curva ROC (Relative Operating Characteristic), donde se indica la precisión para cada una de las clases.

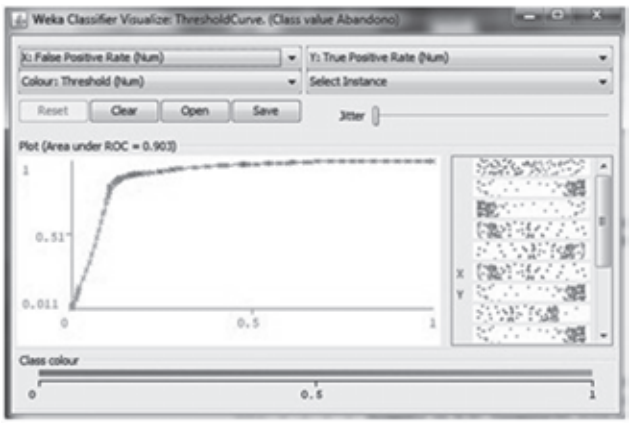

Figura 9: Curva ROC Clase Abandono.

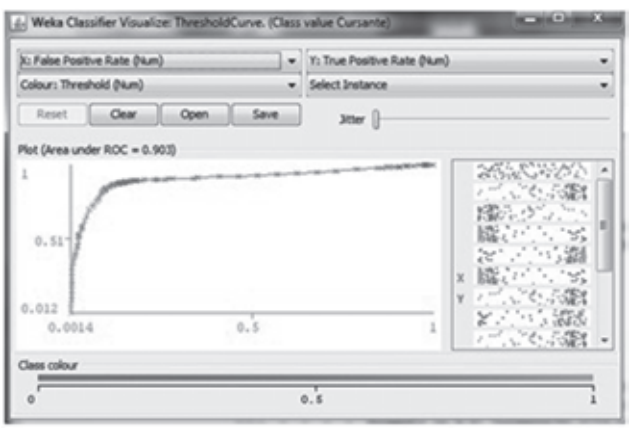

Figura 10 : Curva ROC Clase Cursante.

Queda así de manifiesto la utilidad de la herramienta de evaluación de la calidad del modelo a través de la curva ROC ya que, para el resto de las clases que se pretenden utilizar, la precisión supera el objetivo planteado como requerimiento.

2.3 Características del modelo obtenido

Fase 8: Desarrollo del software predictivo

En esta etapa se han definido las características de los requerimientos necesarios de satisfacer por un software, que implemente el aprendizaje obtenido, mediante las fases metodológicas desarrolladas previamente.

Las reglas de inferencia obtenidas corresponden al algoritmo SimpleCART y se muestra en la siguiente tabla:

```
CART Decision Tree

Edad < 21.5
|  Edad < 20.5: Cursante(551.0/8.0)
|  Edad >= 20.5
|  |  mat_aprob < 18.5: Cursante(278.0/61.0)
|  |  mat_aprob >= 18.5
|  |  |  Promedio < 6.5: Abandono(16.0/2.0)
|  |  |  Promedio >= 6.5
|  |  |  |  mat_aprob < 28.5: Cursante(11.0/1.0)
|  |  |  |  mat_aprob >= 28.5: Abandono(6.0/0.0)
Edad >= 21.5
|  mat_aprob < 0.5
|  |  orientacion_recibida= "Si": Cursante(20.0/0.0)
|  |  orientacion_recibida!= "No": Abandono(817.0/130.0)
|  mat_aprob >= 0.5
|  |  mat_aprob < 11.5: Cursante(643.0/96.0)
|  |  mat_aprob >= 11.5
|  |  |  Promedio < 5.5: Abandono(247.0/37.0)
|  |  |  Promedio >= 5.5
|  |  |  |  mat_aprob < 24.5
|  |  |  |  |  Promedio < 6.5
|  |  |  |  |  |  mat_aprob < 15.5: Cursante(25.0/9.0)
|  |  |  |  |  |  mat_aprob >= 15.5: Abandono(62.0/25.0)
|  |  |  |  |  Promedio >= 6.5: Cursante(51.0/13.0)
|  |  |  |  mat_aprob >= 24.5: Abandono(142.0/25.0)
```

Tabla 10: Resultado del algoritmo SimpleCART.

Esta formulación de requerimientos facilita al desarrollador de software la realización de un programa. Este software será el que ejecute el modelo para nuevos datos proporcionados como entrada.

Conclusiones

3.1 Ventajas de la aplicación metodológica ABI a un sistema informático

Se ha utilizado una base de datos con una importante cantidad de atributos (variables), que han sido objeto para la aplicación de la metodología ABI, concluyendo en un modelo predictivo que facilita el desarrollo de un sistema informático para la detección temprana de alumnos con riesgo de deserción estudiantil.

A través de los resultados obtenidos en las distintas fases, se demuestra que la aplicación estricta de la metodología ABI utilizada en este trabajo, avanza mucho más allá del razonamiento basado en eventos pasados, algunos de ellos provistos por las herramientas típicas de los sistemas de soporte a las decisiones.

El conocimiento obtenido se potencia al aplicar esta metodología, pudiendo responder preguntas de dificultosa resolución o que demandan excesivo tiempo ante tal cantidad de variables.

Es así que a partir de los registros de datos que surgen desde las bases de datos de los estudiantes, ha sido posible encontrar patrones ocultos e información predictiva útil para los decisores.

El modelo logrado satisface los objetivos del proyecto, demostrando así las virtudes de la nueva metodología seleccionada para este trabajo y las capacidades de la aplicación de herramientas como Weka para estos propósitos. [8]

3.2 Detección temprana de alumnos que potencialmente desertan – Lecciones aprendidas

Con relación a las características del conocimiento adquirido, acorde a las reglas de clasificación, se sintetizan algunas de las más relevantes, dejando el resto como base de nuevos estudios para los expertos disciplinares de la educación:

a. Alumnos menores de 21.5 años desertan si transcurridos 2 años su promedio es inferior a 6.5

b. Alumnos mayores de 21.5 años que no han recibido orientación vocacional al ingreso a las carreras y sin asignaturas aprobadas dentro de los 2 años, desertan en un porcentaje de 24.9% (un cuarto del alumnado ingresante).

c. Alumnos mayores de 21.5 años, posiblemente en un segundo año de la carrera (>=11.5 asignaturas aprobadas), cuyo promedio no supera 5.5, abandonan las carreras (>7.5%).

Mediante estos resultados, es posible la aplicación de las reglas obtenidas directamente a la base de datos del SIU Guaraní, facilitando así, la detección de aquellos alumnos que actualmente se encuentran en riesgo de deserción.

Estas conclusiones permitirán la detección temprana de casos de riesgo y de esta manera facilitarán la adopción de las decisiones necesarias para orientar los estudios y transformarlos en exitosos.

Referencias

Arancibia, J.G.: Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM. http://yoshibauco.wordpress.com/ (2011) [1]

SAS Institute. United Kingdom. http://www.sas.com/ - Último acceso: (2011) [2]

Azevedo, A.; Santos, M.F.: KDD, SEMMA y CRISP-DM: A Parallel Overview. IADIS European Conference Data Mining 2008. Part of MCCSIS (2008)[3]

Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T., Shearer, C.; Wirth, R.: CRISP-DM 1.0 Step-by-step data mining guide, CRISP-DM consortium. (1999, 2000)[4]

Cios, K.J.; Pedrycz, W.; Swiniarski, R.W.; Kurgan, L.A.: Data Mining. A Knowledge Discovery Approach. Springer. (2007)[5]

Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R.: Guía paso a paso de Minería de Datos. (2007)[6]

Fayyad, U.; Piatetsky-Shapiro, G.; Smith, P.; Uthurusamy R.: From data mining to knowledge discovery: an overview. In: Advances in Knowledge Discovery and Data Mining. pp. 1-29. California: AAAI Press / The MIT Press. (1996)[7]

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.: The WEKA Data Mining Software: An Update. Pentaho Corporation. (2009)[8]

Watson, H.J.; Wixom, B.H.: The Current State of Business Intelligence. Computer Magazine, Vol. 9 Issue 40, Page(s): 96-99. (2007) [9]

Michalewicz, Z.; Schmidt, M.; Michaelewicz, M.; Chiriac C.: Adaptative Business Intelligence. Springer (2007) [10]

Una Metodología para el Desarrollo de Aplicaciones Autoadaptativas basada en Business Intelligence. Aplicación en Medicina. Tesis para optar a la titulación de postgrado correspondiente a la Maestría en Ingeniería de Software. (2012)[11]

Moss, L.T.; Atre, S. Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications. Addison Wesley. (2003)[12]

Sistema de Autogestión de Alumnos SIU-Guaraní. http://www.siuguarani.com.ar. - Ultimo acceso 6/2012.[13]

La gestión de los sistemas de información en la empresa. Teoría y casos prácticos. Arjonilla Domínguez, S.J.; Medina Garrido, J.A. Tercera Edición. Ediciones Pirámide. Madrid. (2009) [14]