Análisis taxonómico predictivo aplicado a la detección temprana de alumnos universitarios en riesgo de deserción

Marcelo Fabio Roldan[1], Ana Funes[2], Germán Montejano[2]

[1] Universidad Nacional de La Rioja,
Luis M. de la Fuente, 5300 La Rioja, Argentina
Marcelo.roldan@unlar.edu.ar
[2] Universidad Nacional de San Luis,
Ejército de los Andes 950, Argentina
{afunes, gmonte}@unsl.edu.ar

Resumen. La deserción universitaria es un problema que aqueja tanto a la educación pública como privada en la Argentina. En este trabajo se presenta una aplicación autoadaptativa predictiva para la detección temprana de alumnos universitarios que se encuentran en riesgo de abandonar sus estudios.

La aplicación fue construida a partir del modelo obtenido con una herramienta de extracción de conocimiento, sobre datos de alumnos universitarios, aplicando las fases de una metodología de Adaptive Business Intelligence.

Para esto se han tomado en consideración los datos socio-económico-culturales de los alumnos ingresantes así como la finalización de sus estudios, todos obtenidos del Sistema de Gestión Universitaria (SIU). Con estos datos, y aplicando una metodología de Adaptive Business Intelligence se ha generado un modelo de aprendizaje que permite la clasificación de alumnos universitarios como posibles candidatos a la deserción de su estudios.

Palabras claves: Adaptive Business Intelligence, Software Predictivo, Deserción Universitaria, Sistema de Gestión Universitaria, SIU.

1 Introducción

El uso de herramientas estratégicas en el escenario educativo, como la planteada en el presente trabajo, se ha vuelto de fundamental importancia tanto para las universidades públicas como privadas.

La posibilidad de facilitar al decisor información precisa y confiable respecto a indicadores como el porcentaje de alumnos que ha 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.

Por otro lado, existe otro tipo de información estratégica, que no es brindada por una simple herramienta de informes y que puede ser obtenida por técnicas de minería de datos. Es allí donde el aporte de las tecnologías de Adaptive Business Intelligence (ABI) [9] pueden 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 [10] mostrada en la Figura 1.

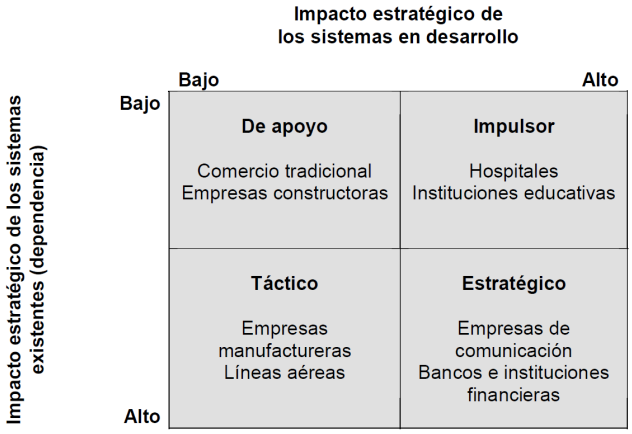

Figura 1. Matriz de McFarlan

La matriz de McFarlan 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 [2].

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.

En este trabajo presentamos la aplicación de una metodología ABI [11] para la extracción de reglas de clasificación de alumnos universitarios, usando como datos de entrada un banco de datos proporcionado por el Sistema Informático Universitario (SIU). Esta base de datos contiene información respecto a los alumnos y considera diferentes aspectos, tales como datos personales, económicos, sociales, culturales; todos ellos favorecen un análisis multivariado. Para este análisis hemos utilizado 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 ha resultado en un árbol de clasificación que modela aquellas causalidades principales de la deserción de los alumnos universitarios, en la institución referida.

La investigación presentada se ha realizado sistemáticamente aplicando paso a paso la metodología ABI tal como lo indica su autor [11]. Los principales resultados obtenidos, al aplicar la metodología

mencionada previamente, han orientado gradualmente el tratamiento de la información oculta en las numerosas variables del caso. Este procesamiento, ha resultado en la determinación de aquellos parámetros cuyo impacto en la deserción es alto.

De manera indirecta, el modelo resultante de este trabajo ha facilitado la construcción de un software que permite el ingreso directo de los datos, indicando de manera predictiva, para cada nuevo caso, si se trata de un alumno potencialmente desertor y cuya directa aplicación a la base de datos del SIU Guaraní, facilitará la detección de aquellos alumnos que se encuentren en riesgo de deserción, con la consecuente posibilidad de adopción de medidas correctivas y de apoyo por parte de la institución.

El resto del trabajo se encuentra organizado de la siguiente manera. La sección 2 presenta los fundamentos teóricos abarcando la tecnología de extracción de conocimiento y la metodología utilizada. La sección 3 describe la procedencia de los datos. En la sección 4 aplicamos la metodología seleccionada al caso de estudio de predicción de deserción de los estudiantes. Finalmente, la sección 5 cierra el trabajo con las conclusiones.

2 Metodología aplicada

El descubrimiento de información “oculta” en grandes volúmenes de datos es posible gracias a técnicas de *minería de datos* (*data mining*) [5], 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) [6] 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* (BI) [8], [9] provee beneficios que se aplican no solamente en los ámbitos empresariales. 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* 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?

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 diversas áreas, 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 [5]. Contar con una metodología, se ha convertido en algo tan importante y necesario como la carta de presentación de las empresas. Con esta premisa en mente, hemos aplicado una nueva metodología para desarrollar un sistema de ABI, a partir del estudio de numerosas variables que tienen correlación en mayor o menor medida con el indicador objetivo estudiado (deserción estudiantil). Dicha metodología [11] surgió a partir de los conceptos relacionados entre disciplinas afines tecnológicamente como OLAP, acceso a datos multiplataformas, Business Intelligence, Data mining y Adaptabilidad, analizándose, para su desarrollo, las interrelaciones existentes entre ellas, y de este modo, conformando un entorno de soporte para las aplicaciones predictivas.

Las etapas de la metodología empleada para el desarrollo de aplicaciones basadas en ABI 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, esta metodología propone el uso de técnicas implementadas de minería de datos, buscando aquellos resultados que proporcionen la información necesaria para acercarse a los objetivos del proyecto. Esto involucra las etapas referidas como de “Búsqueda de patrones, reglas o grupos”, la etapa de “Modelizado predictivo” y la “Validación del modelo”.

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.

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. 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 Figura 2.

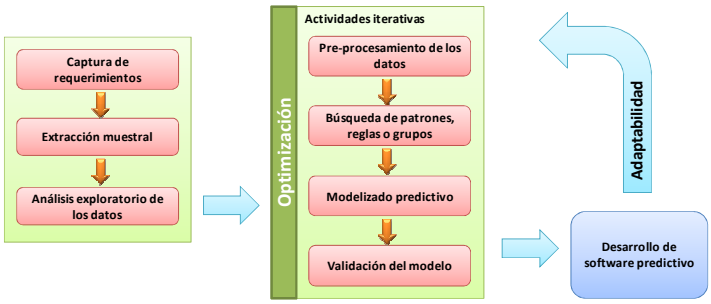

Figura 2. Implementación de la optimización, predicción y adaptabilidad en la metodología.

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 geográfica de sus sedes, 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 SIU (Sistema Informático Universitario), y en particular el SIU Guaraní, usado en este trabajo, 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 seguridad y disponibilidad de la información, fiabilidad, sencillez y costo entre otras ventajas [2].

El SIU Guaraní como sistema de información permite el seguimiento de todas las actividades que realiza un estudiante, desde la inscripción a exámenes y cursadas, 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, hasta la recepción de mensajes [13].

Entre los usos parametrizables que las universidades pueden utilizar, el SIU-Guaraní se presenta como un sistema informático de autogestión académica por Internet, lo que permite a los alumnos un uso más adecuado a 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 requieren usos 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 amplio, el cual puede contener datos de mayor relevancia a diferentes contextos de análisis. En la Figura 3 se expone, a modo de ejemplo, una pantalla donde se resumen los datos económicos de un alumno.

Con los datos recopilados por el SIU Guaraní, en su base de datos, la que aporta datos de característica socio-económico-cultural de cada uno de sus estudiantes, hemos construido nuestra referencia inicial para el procesamiento de los datos mediante herramientas ETL (Extract-Transform-Load) los que constituirán los datos de entrada para el análisis de patrones.

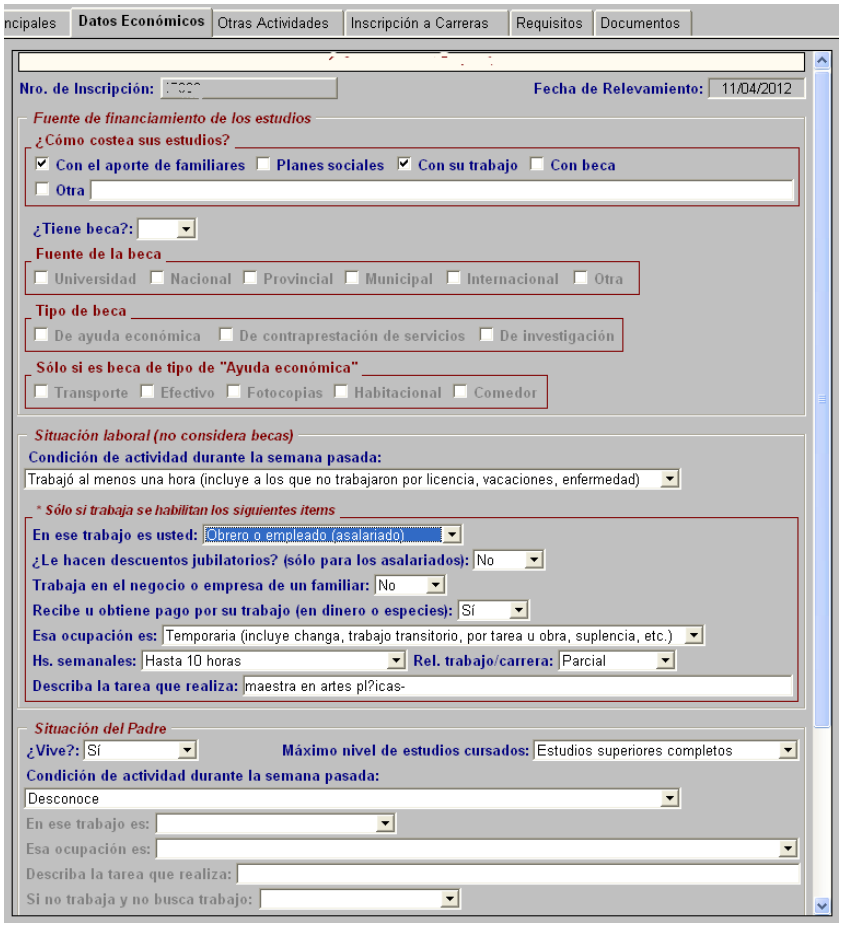

Figura 3. Ficha de datos económicos del alumno. Interface con el usuario del SIU Guaraní

4 Aplicación de la metodología para la predicción de potenciales casos de deserción

Los datos empleados en el análisis provienen de la base de datos del SIU Guiaraní. En ellos se encuentran alumnos en diferentes estadios de sus estudios con características sociales diversas. Se ha abordado el problema utilizando Weka [7] [14] como herramienta automatizada.

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 pudiera ser relevante, dejando de lado algunos como los 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 definió como requisito la construcción de un modelo que realice la clasificación de alumnos con riesgo de deserción con una precisión cuyo error absoluto medio sea <=2% y su precisión de al menos 80%, proveyéndose, al finalizar el proyecto, el conjunto de 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.

Aspectos relevantes del proyecto

Tabla 1. Aspectos relevantes

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

Fase 2: Extracción muestral

Se trabajó con un banco de datos con un elevado número de muestras (4707 registros).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.

Fase 3: Análisis exploratorio de los datos

Información de los atributos Distribución de clases

Las cantidades de muestras por cada clase se pueden observar en la Tabla 2 , donde la frecuencia para la cantidad de alumnos que han abandonado asciende a 2456, mientras que los que se mantienen cursando las carreras es de 2251 alumnos.

Tabla 2. Distribución por clase

Clase	Número de muestras
Abandono	2456
Cursante	2251

Ya que este tipo de sistemas son conducidos por datos, es importante tener una buena comprensión de los mismos (Tabla 3).

Tabla 3. Detalle de los atributos

#	Fieldname	Type	Length	Precision	Step origin	Storage	Mask
1	nombre_alumno	String	-	-	Datos de alumnos	normal	
2	nro_documento	Number	-	-	Datos de alumnos	normal	#
3	fecha_inscripcion	Date	-	-	Datos de alumnos	normal	yyyy/MM/dd
4	nacionalidad	String	-	-	Datos de alumnos	normal	
5	fecha_nacimiento	Date	-	-	Datos de alumnos	normal	yyyy/MM/dd
6	Edad	Integer	-	0	Datos de alumnos	normal	#
7	localidad_nacimiento	String	10	-	Datos de alumnos	normal	
8	provincia_nacimiento	String	10	-	Datos de alumnos	normal	
9	colegio_secundario	String	10	-	Datos de alumnos	normal	
10	titulo_secundario	String	10	-	Datos de alumnos	normal	
11	orientacion_recibida	String	-	-	Datos de alumnos	normal	
12	estado_civil	String	-	-	Datos de alumnos	normal	
13	vive_unido_de_hecho	String	-	-	Datos de alumnos	normal	
14	cant_hijos	Integer	-	0	Datos de alumnos	normal	#
15	obra_social	String	16	-	Datos de alumnos	normal	
16	residencia_tipo	String	-	-	Datos de alumnos	normal	
17	con_quien_vive	String	10	-	Datos de alumnos	normal	
18	costea_estudios	String	20	-	Datos de alumnos	normal	
19	tiene_beca	String	-	-	Datos de alumnos	normal	
20	situacion_laboral	String	-	-	Datos de alumnos	normal	
21	padre_vive	String	-	-	Datos de alumnos	normal	
22	max_est_cur_padre	String	-	-	Datos de alumnos	normal	
23	madre_vive	String	-	-	Datos de alumnos	normal	
24	max_est_cur_madre	String	-	-	Datos de alumnos	normal	
25	DISP__PC_EN_CASA	String	-	-	Datos de alumnos	normal	
26	DISP__PC_EN_TRABAJO	String	-	-	Datos de alumnos	normal	
27	DISP__PC_EN_UNIVERSIDAD	String	-	-	Datos de alumnos	normal	
28	DISP__PC_EN_OTRO_LUGAR	String	-	-	Datos de alumnos	normal	
29	accede_internet_casa	String	-	-	Datos de alumnos	normal	
30	accede_internet_trabajo	String	-	-	Datos de alumnos	normal	
31	accede_internet_universidad	String	-	-	Datos de alumnos	normal	
32	accede_internet_cyber	String	-	-	Datos de alumnos	normal	
33	accede_internet_otro_lugar	String	-	-	Datos de alumnos	normal	
34	regularidad_accede_a_internet	String	-	-	Datos de alumnos	normal	
35	practica_deportes	String	-	-	Datos de alumnos	normal	
36	habla_ingles	String	-	-	Datos de alumnos	normal	
37	nombre_carrera	String	-	-	Datos de alumnos	normal	
38	sexo	Integer	-	0	Datos de alumnos	normal	#
39	carrera_cod	String	-	-	Datos de alumnos	normal	
40	plan	Integer	-	0	Datos de alumnos	normal	#
41	legajo	String	-	-	Datos de alumnos	normal	
42	Promedio	Number	-	-	Datos de alumnos	normal	#,#
43	mat_aprob	Integer	-	0	Datos de alumnos	normal	#
44	fech_prim_exa	String	-	-	Datos de alumnos	normal	yyyy/MM/dd
45	fech_ult_exa	String	-	-	Datos de alumnos	normal	yyyy/MM/dd
46	longevidad_alumno	Integer	-	0	Datos de alumnos	normal	#
47	class	String	-	-	Datos de alumnos	normal	

Uno de los objetivos de esta etapa es 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 4).

En esta etapa, la exploración de los datos 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 pueden tomar todas las variables/características que se necesitan 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.

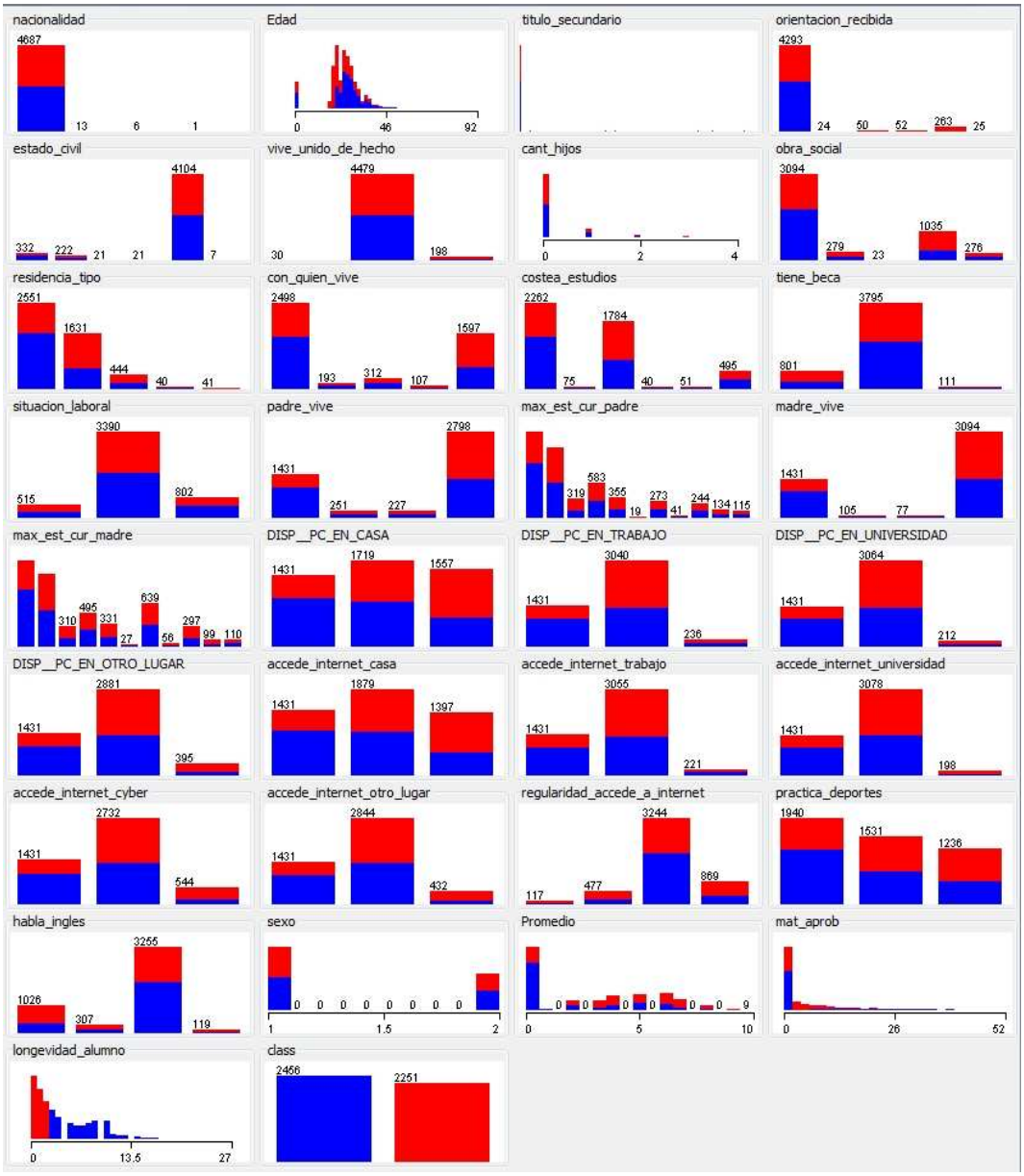

Figura 4. Relaciones entre los datos

Fase 4: Pre-procesamiento y tratamiento de los datos

Para la realización del experimento, se han revisado los datos encontrándose las relaciones entre ellos mostradas en la Figura 4. La distribución de algunos datos, ha indicado la posibilidad de realizar una selección de un subconjunto de datos, basados en algunas correlaciones posibles de estimar previamente a la modelización.

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.

Titulo 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.

Existen un grupo de 1431 alumnos que no han respondido a la mayoría de los parámetros, entre algunos de los mencionados anteriormente, por lo que se los ha filtrado 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 pre-procesamiento:

Tabla 4. Detalle de los atributos que requirieron pre-procesamiento

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

Los datos procesados se pueden visualizar en la Figura 5, donde se observa una distribución con mejores características que los de la Figura 4.

Esto ha repercutido en acciones sobre el resto de los atributos, como se ve en la Tabla 5.

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 en la Tabla 6.

El análisis utilizando el algoritmo de filtrado InfoGainAttributeEval [9] a través de un método de ranking permitió clasificar el ranking de atributos obteniendo las sensibilidades del Subset Nº 4 denominado InfoGain Attribute Evaluation.

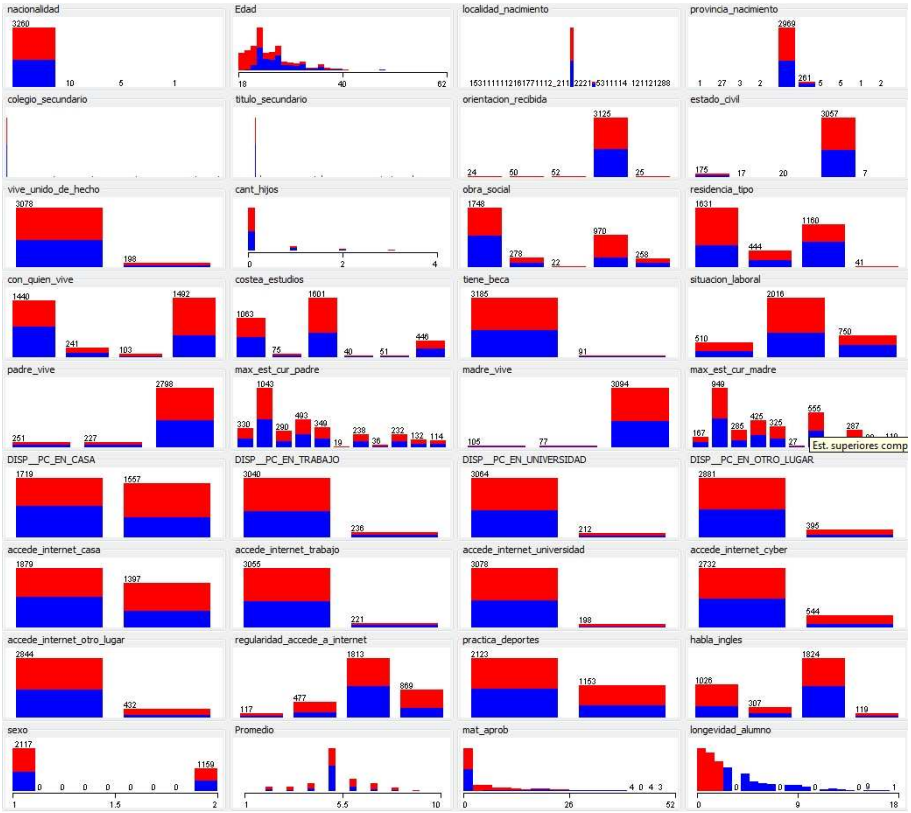

Figura 5. Relaciones entre los datos procesados

Tabla 5. Acciones realizadas sobre los atributos

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

Tabla 6. Filtros de selección de atributos

Subset Nº	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 ...

Fase 5: Búsqueda de patrones, reglas o grupos

Se aplicaron los algoritmos de clasificación en árbol SimpleCart, J48 y Decision Table, los que proveen un método supervisado para la clasificación. Los resultados obtenidos se muestran en la Tabla 7.

Tabla 7. Resultados obtenidos en la clasificación de los subconjuntos de datos

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 %

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 método de ranking [9].

Fase 6: Modelado predictivo

Para el proceso de entrenamiento del modelo, en Weka, utilizamos validación cruzada junto a 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”, para eliminar los atributos con menor significancia. Luego de ello, el porcentaje final de acierto asumido como el mejor es el mostrado en la Tabla 8.

Tabla 8. Resultado del modelado predictivo

```
=== 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
```

Fase 7: Validación comparativa del modelo

La Figura 6 muestra las evaluaciones del modelo realizadas a través del área bajo la curva ROC, donde se indica la precisión para cada una de las clases.

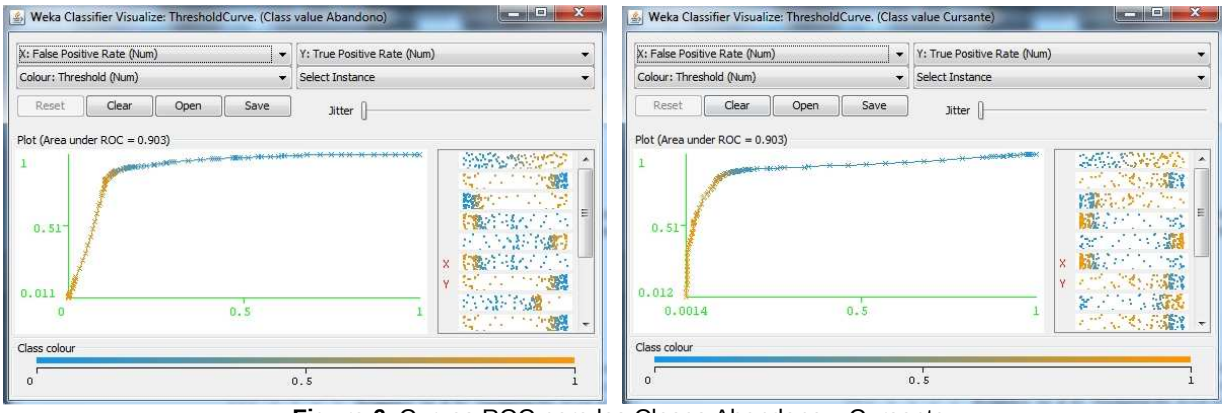

Figura 6. Curvas ROC para las Clases Abandono y Cursante

Fase 8: Desarrollo del software predictivo

En esta etapa se desarrolló el software que implementa el modelo aprendido mediante las fases metodológicas ABI y que se reducen a la implementación de las reglas de inferencia relacionadas con las principales variables que han sido priorizadas como atributos de mayor incidencia en las causas de deserción (Edad, orientacion_recibida, mat_aprob y longevidad_alumno). Dichas reglas corresponden al algoritmo SimpleCART y se muestran en la Tabla 9.

Tabla 9. Resultado del algoritmo SimpleCART

```
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)
```

Este software será el que ejecute el modelo para nuevos datos proporcionados como entrada, permitiendo clasificar nuevos individuos como potenciales desertores.

5 Conclusiones

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:

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

Asimismo, se ha utilizado una base de datos con una importante cantidad de atributos, 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 ha demostrado que la aplicación estricta de la metodología ABI utilizada en este trabajo, permite avanzar 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 proceso metodológico ha resultado exitoso, obteniéndose un modelo de manera natural, permitiendo autoevaluar los resultados alcanzados mediante la aplicación de las técnicas y algoritmos de Data mining en el contexto de ABI.

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.

Referencias

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

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

[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]. Nolan, R. and McFarlan, F. W.: Information Technology and the Board of Directors, in the Harvard Business Review (2005).

[11]. Roldán, Marcelo. Una Metodología para el Desarrollo de Aplicaciones Autoadaptativas basada en Business Intelligence. Aplicación en Medicina. Tesis para optar a la titulo de Magister en Ingeniería de Software, Universidad Nacional de San Luis. (2012).

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

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

[14]. The University of Waikato, Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/index.html, Ultimo acceso 27/04/2013.