Análisis de rendimiento académico estudiantil usando data warehouse y redes neuronales

Analysis of students' academic performance using data warehouse and neural networks

Carolina Zambrano Matamala[1] Darío Rojas Díaz[1] Karina Carvajal Cuello[1]
Gonzalo Acuña Leiva[2]

Recibido 12 de agosto de 2011, aceptado 9 de diciembre de 2011
Received: August 12, 2011 Accepted: December 9, 2011

RESUMEN

Cada día las organizaciones tienen más información porque sus sistemas producen una gran cantidad de operaciones diarias que se almacenan en bases de datos transaccionales. Con el fin de analizar esta información histórica, una alternativa interesante es implementar un Data Warehouse. Por otro lado, los Data Warehouse no son capaces de realizar un análisis predictivo por sí mismos, pero las técnicas de inteligencia de máquinas se pueden utilizar para clasificar, agrupar y predecir en base a información histórica con el fin de mejorar la calidad del análisis. En este trabajo se describe una arquitectura de Data Warehouse con el fin de realizar un análisis del desempeño académico de los estudiantes. El Data Warehouse es utilizado como entrada de una arquitectura de red neuronal con tal de analizar la información histórica y de tendencia en el tiempo. Los resultados muestran la viabilidad de utilizar un Data Warehouse para el análisis de rendimiento académico y la posibilidad de predecir el número de asignaturas aprobadas por los estudiantes usando solamente su propia información histórica.

Palabras clave: Data warehouse, análisis histórico, predicción, redes neuronales, información estratégica.

ABSTRACT

Every day organizations have more information because their systems produce a large amount of daily operations which are stored in transactional databases. In order to analyze this historical information, an interesting alternative is to implement a Data Warehouse. In the other hand, Data Warehouses are not able to perform predictive analysis for themselves, but machine learning techniques can be used to classify, grouping and predict historical information in order to improve the quality of analysis. This paper depicts architecture of a Data Warehouse useful to perform an analysis of students' academic performance. The Data Warehouse is used as input of a Neural Network in order to analyze historical information and forecast. The results show the viability of using Data Warehouse for academic performance analysis and the feasibility of predicting the number of approved courses for students using only their own historical information.

Keywords: Data warehouse, neural networks, historical analysis, prediction, strategic information.

1 Departamento de Ingeniería Informática y Ciencias de la Computación. Universidad de Atacama. Avenida Copayapu 485. Copiapó, Chile. E-mail: carolinazambrano@gmail.com; dfrojas@gmail.com; karina.carvajal@uda.cl.
2 Departamento de Ingeniería Informática. Universidad Santiago de Chile, Avenida Ecuador 3659, Estación Central, Santiago, Chile. E-mail: gacuna@usach.cl

INTRODUCCIÓN

Una de las acciones más utilizadas en las instituciones educacionales para dar valor a la información y dar apoyo a la toma de decisiones, es la confección de reportes. La confección de los reportes es una acción exploratoria, es decir, se hacen ciertos cruces de datos y, dependiendo de los resultados, se van analizando otros criterios hasta que se llega a un punto en el cual los resultados son satisfactorios para tomar decisiones sobre la organización. El apoyo a la toma de decisiones puede ser realizado mediante sistemas especialmente diseñados para ello como son los DSS [21] (Decision Support Systems), los cuales pueden generar informes parametrizables en forma periódica, rápida y fácil, como los presentados en [17].

Otro método comúnmente utilizado es la creación de reportes mediante la manipulación directa de bases de datos transaccionales a través del lenguaje SQL (Structured Query Language), lo cual tiene el inconveniente de requerir una persona experta en la utilización de SQL. Además el desarrollo de reportes puede tomar un tiempo considerable debido a que las bases de datos transaccionales no están diseñadas específicamente para el análisis. Otro método muy utilizado trata sobre el uso de planillas de cálculo y datos tabulados; sin embargo, este método a pesar de necesitar menos conocimientos técnicos sufre de la imposibilidad de manejar eficientemente grandes cantidades de datos directamente, como también sufren de la dificultad de poder realizar el cruzamiento de datos en forma sencilla desde distintas fuentes de datos.

Por otro lado, los Data Warehouse (DW) son repositorios de datos electrónicos especialmente diseñados para la generación de reportes y análisis de datos [13, 23]. Las características distintivas de los DW respecto a los sistemas descritos anteriormente es que son flexibles, integran todos los aspectos organizacionales de interés, pueden manejar grandes volúmenes de datos eficientemente, permiten la creación y cálculo de indicadores de gestión. Además, los DW se diseñan con el objetivo de ser eficientes en los requerimientos de análisis para niveles estratégicos en las organizaciones, por lo que toman en cuenta los objetivos estratégicos de la organización directamente [15]. En el mismo contexto, los DW permiten analizar de forma eficiente la información histórica de una organización, y de esta forma visualizar tendencias de comportamiento de los indicadores de gestión en el tiempo. Sin embargo, a pesar de que la información histórica nos puede dar un indicio de la tendencia histórica que puede seguir un indicador, no es suficiente para predecir con certeza algún indicador en particular.

Sin embargo, un DW sí puede proveer de una base sólida de análisis y comportamiento inicial o de entrada para técnicas de Inteligencia de Máquinas [16] que permitan aprender los patrones de estos indicadores para poder predecir patrones futuros. Para esto último, las Redes Neuronales Artificiales (RNA) son algoritmos que tienen la capacidad de asociar o clasificar patrones, comprimir datos, controlar procesos y aproximar funciones no lineales [9, 11].

Las RNA son estructuras simplificadas de lo que se conoce acerca de los mecanismos y estructura física del conocimiento y aprendizaje biológico, tomando como base el funcionamiento de la neurona biológica. Una RNA es una estructura paralela de procesamiento distribuido de la información, cuyo elemento básico es la neurona [9, 11]. Existen distintos tipos de redes neuronales, dependiendo del tipo de aprendizaje que se desee realizar. El tipo de red neuronal más utilizado en clasificación y predicción es el Perceptrón Multicapas, que consiste de neuronas conectadas por capas, donde cada capa tiene una cantidad de neuronas asociadas. El aprendizaje que se utiliza en este tipo de redes es el de retropropagación del error, en donde se trata de minimizar la función del error entre la salida deseada y la del modelo neuronal a partir de un conjunto de observaciones ya clasificadas [9, 11].

Las RNA han sido ampliamente utilizadas en el contexto de predicción de sistemas complejos. En efecto, es sabido que predictores autorrecurrentes con o sin entrada exógena NAR o NARX pueden ser fácilmente aproximados mediante redes neuronales. Diaconescu [6] lo hace en el caso de predicción de series de tiempo caóticas, y Jiang and Song realizan lo propio para predecir el comportamiento de series de datos financieros [12]. Incluso predictores más sofisticados, como aquellos autorrecurrentes y que consideran errores de predicción anteriores como parte de su regresor (NARMA o NARMAX), han

sido exitosamente aproximados mediante redes neuronales combinadas con lógica difusa [7].

En el contexto del uso de RNA para predecir patrones futuros de comportamiento en el ámbito educativo, trabajos como [19] utilizan una red neuronal multicapas para predecir el éxito o fracaso de estudiantes utilizando los datos de PISA, obteniendo una precisión de más del 75% en la clasificación. Por otro lado, en [4] se utiliza una red neuronal para predecir el rendimiento en la asignatura de Algoritmos y Programación I; para ello utilizan los datos de 450 estudiantes, en dos redes neuronales, una para pronóstico (aprueba o no) y otra para guiar en temas de estudios para aprobar. En [20] utilizan una red neuronal multicapas para predecir el rendimiento de estudiantes de primer año de la carrera de Ingeniería Civil en la Universidad de Concepción. La estructura de la red neuronal propuesta tiene una precisión que fue cercana al 91%, mostrando que las variables más importantes (las de mayor incidencia en una correcta decisión) para esta experiencia fueron el sexo, puntaje de ingreso, estrato socioeconómico y la distancia entre la residencia y la universidad, sin embargo, el puntaje de ingreso consigue imponerse por sobre las otras características. Estos trabajos descritos están preocupados de intentar predecir el éxito o fracaso de los estudiantes en primer año, o su inserción o deserción de la universidad, tomando siempre como objetivo predecir valores nominales y atemporales. En este mismo contexto, el enfoque del presente trabajo es realizar un análisis temporal del desempeño de los estudiantes, considerando para ello la capacidad de predecir el comportamiento futuro de un alumno en cualquier punto de avance en su desarrollo dentro de la carrera. Por ejemplo, el enfoque propuesto permite predecir la cantidad de asignaturas que un alumno tomará en un semestre y cuántas de ellas aprobará, sólo considerando los datos del currículo del semestre anterior y las condiciones de entrada.

En este trabajo se ha implementado un DW en base a información obtenida de un sistema de base de datos no relacional (basado en archivos o también llamado sistema heredado). El DW se ha diseñado para el análisis del comportamiento de aprobación y avance en una malla curricular con datos reales de los currículos de los estudiantes de la carrera de Ingeniería Civil en Computación e Informática de la Universidad de Atacama. El DW no está enfocado sólo en el análisis de comportamientos históricos de los estudiantes, sino que también ha sido pensado como una arquitectura base para la predicción de tendencias futuras a través de técnicas de RNA. Es importante indicar que este trabajo es una extensión a las investigaciones presentadas en [24, 25, 27].

El artículo tiene la siguiente estructura: primero se presenta un apartado de Metodología que expone la metodología de trabajo que explica la arquitectura del DW implementado y la arquitectura de la RNA utilizada. Posteriormente se presenta un apartado de Análisis y Resultados que incluye los principales resultados del análisis ROLAP obteniendo las tendencias de comportamiento, para luego presentar los resultados de la predicción de rendimiento en base a la RNA. Finalmente se presentan la Conclusión y Trabajos Futuros que incluyen comentarios sobre los resultados y potenciales tareas que restan por desarrollar.

METODOLOGÍA

Implementación del Data Warehouse

Un DW está compuesto de elementos básicos, entre los que podemos encontrar las dimensiones de análisis, las medidas también conocidas como indicadores de gestión y los hechos que representan los datos reales. En este contexto, los DW se diseñan para poder calcular y analizar un conjunto de indicadores de gestión. Con este enfoque, los "indicadores de gestión" dirigirán el diseño, y se convertirán en las "medidas", y las "variables/criterios" a analizar se convertirán en las "dimensiones" de un modelo multidimensional [13, 23]. Cada celda o hecho contiene uno o más indicadores de gestión, como por ejemplo podría ser la cantidad de estudiantes por asignatura y región, promedio de notas, etc. Otro concepto en el ámbito de los DW es el de Data Mart que representa pequeños DW centrados en un tema o un área de negocio específico dentro de una organización [1].

La tecnología que permite una acción exploratoria de los datos del DW se realiza mediante OLAP [5] (Online Analytical Processing), que no sólo permite flexibilidad en cuanto a la navegación a través del modelo multidimensional de la información, sino que también es flexible en la definición de los reportes y aplicaciones que se construyen a partir de ella. Además, las herramientas OLAP definen claramente

operadores especiales de refinamiento o manipulación de consultas que pueden ser comprendidas mucho más fácilmente que las sentencias SQL y que además son eficientes, ya que se realizan sobre datos y resúmenes precomputados.

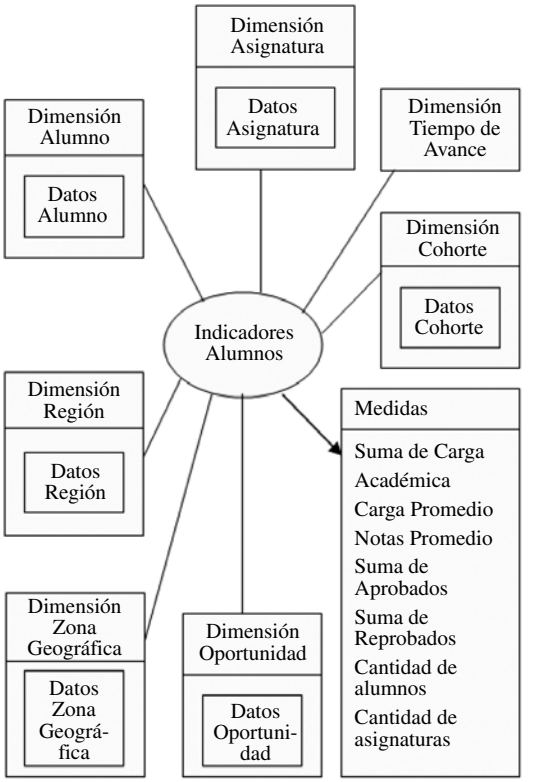

Figura 1. Esquema conceptual del DW (usando modelo conceptual CMDM [3] para especificar el diseño del DW implementado para el análisis de indicadores de estudiantes.

Un sistema de DW puede ser implementado bajo enfoque Molap (MultidimensionalOlap), Rolap (RelacionalOlap) o mediante el híbrido Holap (permite tanto Molap como Rolap) [5]. En este trabajo se utilizó enfoque Rolap. Independiente del enfoque, los principales procesos que se llevan a cabo en el desarrollo de un DW son los siguientes:

- **Proceso de modelamiento conceptual:** El modelo conceptual es independiente de la tecnología y es primordial para especificar los requerimientos de análisis y disponibilidad de información. A nivel de modelos conceptuales de DW no existe consenso en la comunidad de investigadores sobre cuál es el modelo aceptado como estándar para la representación de un DW; sin embargo, hay varias propuestas, algunas de ellas se presentan en [3, 8, 10, 22]. Durante el proceso de modelamiento conceptual se genera el esquema conceptual del DW. En este trabajo se utilizó el modelo conceptual CMDM [3] debido a la sencillez de su notación y porque su objetivo es justamente la especificación conceptual de un DW.
- **Proceso de modelado lógico e implementación Física:** El modelo lógico especifica formalmente el esquema multidimensional, sus restricciones y capacidades. Por otro lado, el esquema lógico es implementado directamente en un motor de base de datos, transformándose en tablas físicas. En el caso de los DW esquemas de diseño lógico son el esquema estrella y el esquema copo de nieve [2]. En la etapa de implementación física se crean las tablas de dimensión y tabla de hecho, dependiendo del tipo de esquema estrella o copo de nieve.
- **Proceso de carga de datos ETL:** El proceso ETL (Extraction, Transformation, Load) es el encargado de extraer los datos de las bases de datos originales, transformarlos y cargarlos en el DW. La Figura 2 muestra un esquema del proceso ETL que se llevó a cabo durante el desarrollo de este trabajo.

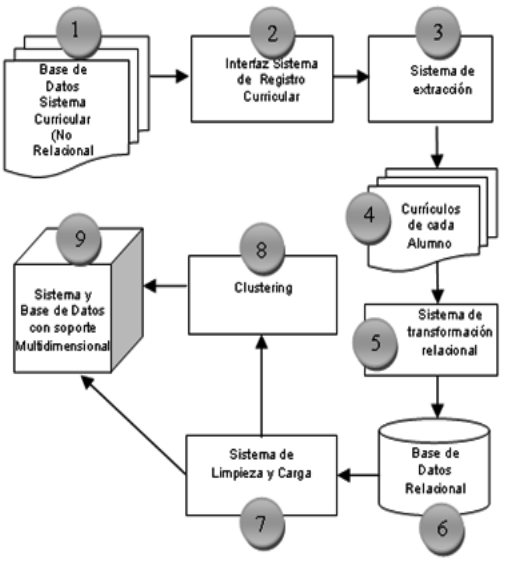

Figura 2. Esquema ETL simplificado para la carga del DW.

- **Proceso de análisis Rolap:** Permite la acción exploratoria a través de las operaciones definidas en Olap para el análisis y creación de reportes bajo modelo relacional.

En la implementación del DW la primera etapa consistió en diseñar el esquema conceptual para el análisis como se muestra en la Figura 1. Este esquema de modelo conceptual posee siete dimensiones de análisis:

1) Alumno: con los datos personales de los estudiantes y su estado.
2) Asignatura: con los datos de las asignaturas impartidas y condiciones de entrada a la universidad (PSU).
3) Región: con las regiones y ciudades de donde provienen los estudiantes.
4) Oportunidad: Representa los datos sobre las oportunidades posibles de cursar las asignaturas.
5) Tiempo de Avance: Tiempo de permanencia de un alumno en la carrera, en base a los semestres.
6) Zona Geográfica: Representa la zona geográfica donde se ubica el alumno.
7) Cohorte: Cohorte a la que pertenecen los estudiantes.

Por otro lado, los indicadores multidimensionales son implementados a través de las Medidas, tales como cantidad de estudiantes, suma de aprobados, etc., según se aprecia en la Figura 1. En este contexto, cabe notar que el esquema lógico no es presentado por efectos de simplicidad y extensión.

El proceso ETL simplificado se presenta en la Figura 2, el cual consiste en extraer los datos desde la base de datos del sistema de información curricular de estudiantes de la universidad (1), el cual no está soportado por un motor relacional y funciona a través de archivos (sistemas heredados). Este sistema sólo es accesible mediante una interfaz de usuario a través de la red mediante una aplicación de consola heredada del lenguaje COBOL (2). Para extraer esta información se simuló el proceso manual de extracción mediante una aplicación especialmente diseñada para ello (3), luego de lo cual se extrajo el currículo de cada alumno en formato de texto (4). Estos archivos de texto son transformados mediante la utilización de un software diseñado a medida (5) y cargados en una base de datos relacional (6), tras lo cual son transformados nuevamente por otra aplicación implementada a través de procedimientos almacenados (7) que los carga en el DW (9).

Por otro lado, los datos escritos en lenguaje natural sobre la dirección particular de los estudiantes son procesados para obtener la ubicación geográfica de los alumnos en coordenadas de latitud y longitud mediante un software de geolocalización especialmente diseñado y basado en la base de datos y API de Google Maps (ver Figura 3a). Con tal de determinar agrupaciones de alumnos automáticamente, dependientes sólo de su ubicación geográfica (única información en este contexto obtenible desde la base de datos original), se realizó un proceso de clustering (8) mediante el algoritmo k-means (14). Este algoritmo consiste en determinar k particiones a partir de n datos (alumnos), donde cada partición o agrupación está definida por un centroide y los elementos que pertenecen a cada agrupación son definidos por un criterio de cercanía o distancia. En este trabajo se estableció $k = 4$ agrupaciones mediante el índice Index I (14), y la distancia utilizada fue la euclidiana. La principal ventaja que supone la agrupación de alumnos según su ubicación a través de técnicas de clustering es que permite determinar agrupaciones en base a las propias características de los elementos, siendo esta selección independiente del criterio humano y por lo tanto más objetiva al momento de utilizarla en el análisis, sin dejar de lado que puede procesar gran cantidad de datos en forma automática y en poco tiempo. La Figura 3b muestra el resultado del proceso de clustering para el caso particular de los alumnos de la carrera de Ingeniería Informática y Ciencias de la Computación, donde se pueden apreciar las cuatro zonas de agrupaciones determinadas. Luego estos datos son cargados igualmente al DW como parte de la dimensión Zona Geográfica, tras lo cual pueden ser procesados por las operaciones comunes del DW.

Implementación de la arquitectura de RNA

En esta etapa se crea la arquitectura de RNA que se alimentará de algunos de los datos obtenidos por medio del DW. Tras obtener el DW cargado (etapa 9 de la Figura 2), se diseñó una arquitectura de RNA para la predicción de rendimiento de los estudiantes usando los algoritmos de Matlab. En este caso la RNA fue utilizada para estimar el comportamiento de un estudiante en el siguiente semestre. La Figura 4 muestra el esquema utilizado.

Figura 3a. Sistema de geolocalización automática utilizando Google Maps como base de datos. (Disponible en http://frodo.diicc.uda.cl/demogeoloc/)

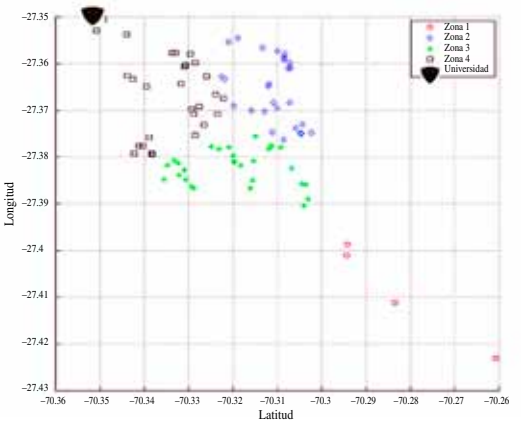

Figura 3b. Resultado de la etiquetación de la ubicación geográfica de los estudiantes mediante clustering.

Las entradas consideradas fueron: puntaje PSU Lenguaje, puntaje PSU Matemáticas, puntaje PSU Historia, puntaje PSU de ingreso, semestre inicial, cantidad de asignaturas inscritas al inicio del semestre y cantidad de asignaturas aprobadas en el semestre. La salida corresponde a la cantidad de asignaturas inscritas y la cantidad de asignaturas aprobadas en el siguiente semestre. Cabe hacer notar que tanto las entradas como las salidas están normalizadas entre los valores 0 y 1.

Se tienen a disposición 4.042 datos temporales, de los cuales se utilizaron 2.515 para entrenamiento, 988 para validación y 563 para prueba. Se realizaron varios entrenamientos cambiando la cantidad de neuronas en la capa oculta, obteniendo como resultado final una red neuronal con la estructura de 10 neuronas de entrada, 8 neuronas en la capa oculta y dos neuronas para la salida. La red neuronal fue entrenada con el algoritmo backpropagation y se utilizó la función logarítmica sigmoide en ambas capas de la red [11].

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n} o_i^2}} \qquad (1)$$

$$RSD = \sqrt{\frac{\sum_{i=1}^{n}(o_i - p_i)^2}{N}} \qquad (2)$$

$$IA = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}\left(|o_i'| + |p_i'|\right)^2} \qquad (3)$$

Los resultados obtenidos se validaron utilizando medidas de desempeño que permiten indicar el grado de generalización del modelo utilizado. Dentro de los índices que se utilizaron se encuentran [9]: el Error Cuadrático Medio (RMS), el Error Residual Estándar (RSD) y el Índice de Adecuación (IA), que se muestran en la ecuaciones (1), (2) y (3), respectivamente, donde *oi* y *pi* son los valores observados y predichos respectivamente en el tiempo i, y N es el número total de datos. Además, *pi*' = *pi* - *om* y *oi*' = *oi* - *om*, siendo *om* el valor medio de las observaciones.

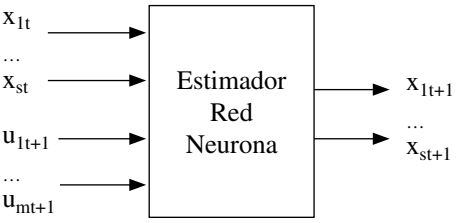

Figura 4. Modelo neuronal de estimación del comportamiento de estudiantes.

El IA indica el grado de ajuste que tienen los valores estimados con los valores reales de una variable; un valor cercano a 1 indica una buena estimación.

Por otro lado, RMS y RSD cercanos a cero indican una buena calidad de ajuste.

ANÁLISIS Y RESULTADOS

En este apartado se analiza el comportamiento de ciertos indicadores en el tiempo a través de la arquitectura de DW implementada y la predicción de alguno de estos indicadores mediante una RNA.

Análisis mediante DW

La plataforma utilizada para el análisis Rolap fue Pentaho Business Intelligence [18], en su versión open source, que cubre las necesidades de Análisis de los Datos y de Reportes, siendo una de sus características su funcionalidad y simplicidad en la implantación.

El objetivo de los análisis que se presentan a continuación es demostrar la versatilidad de los resultados de las operaciones mediante Rolap, debido a que todos los reportes presentados en este trabajo fueron generados en poco tiempo (en relación al diseño e implementación del DW), lo que indica claramente la capacidad de la plataforma DW- Rolap para consultar y analizar datos dispuestos multidimensionalmente desde distintos puntos de vista, sin un diseño preestablecido del sistema, sino más bien sólo del modelo de datos y análisis previo que permite llegar a una arquitectura de diseño de DW robusta para el análisis.

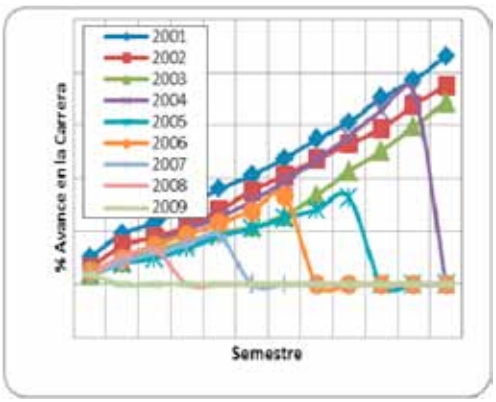

Figura 5. Porcentaje de Asignaturas Aprobadas Acumuladas (PAAA) de la carrera por semestre de permanencia para cada cohorte.

En el gráfico de la Figura 5 se muestra la tendencia del Porcentaje de Aprobación de Asignaturas Acumuladas (PAAA) por semestre de permanencia para las distintas cohortes a partir del año 2001. Como se puede apreciar, en el semestre 12 de permanencia los estudiantes de la cohorte 2001 presentan en promedio un 85% de los ramos de la carrera aprobado, siendo el mejor desempeño según las cohortes analizadas. Por otro lado, se puede ver que las cohortes 2002, 2004 y 2006 se escapan al comportamiento común de las cohortes 2003, 2005, 2007, las cuales tienen un PAAA en el tiempo bastante más bajo. Cabe notar que cohortes más nuevas no poseen más información, debido a que aún no había datos para los semestres posteriores; sin embargo, la tendencia inicial de las curvas permite predecir a simple vista su comportamiento futuro. Cabe destacar que esta predicción es sólo por análisis de la curva del gráfico.

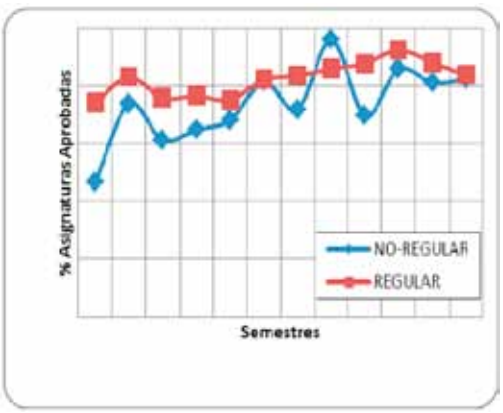

Figura 6. Porcentaje de Aprobación por Semestre (PAS) en asignaturas de la carrera por semestre de permanencia para estudiantes regulares y no regulares.

En el gráfico de la Figura 6 se muestra el Porcentaje de Aprobación de Asignaturas por Semestre (PAS) de los estudiantes regulares de la carrera y los en situación no regular. Los estudiantes no regulares son aquellos estudiantes eliminados o que no renovaron matrículas o que se encuentran en cualquier otra situación que les quite la condición de alumno regular. Como se aprecia, la aprobación de los estudiantes regulares es siempre superior que los estudiantes no regulares, salvo para el semestre 8, el cual presenta una inferioridad respecto a los no regulares. Esto último es debido a que en el semestre 8, hay muy pocos estudiantes en condición no regular y por lo

tanto pocas asignaturas reprobadas en relación a las aprobadas por parte de estos estudiantes.

En el gráfico de la Figura 7 se puede apreciar que el PAAA por zona geográfica es muy similar, excepto para la Zona 1, lo que puede estar justificado por la distancia geográfica de estos pocos estudiantes con respecto a la universidad, la cual se encuentra marcada por su escudo (imagen del escudo UDA) que se observó en la Figura 3b donde se mostró el resultado de la etiquetación de la ubicación geográfica de los estudiantes mediante clustering.

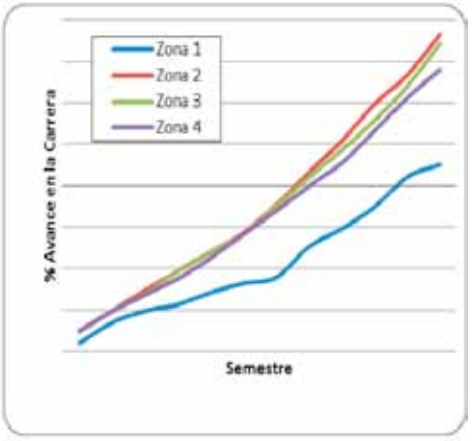

Figura 7. Porcentaje de Asignaturas Aprobadas Acumulada (PAAA) de la carrera por semestre de permanencia para cada zona geográfica.

En el gráfico de la Figura 8 se puede apreciar que las cohortes tienen en general un comportamiento irregular del PAS. Por ejemplo, la cohorte 2002 tiene un comportamiento inferior en porcentaje a las otras cohortes en cada semestre, y además su variabilidad en el tiempo también es mayor. Este comportamiento puede ser explicado porque los estudiantes en un semestre determinado en su mayoría reprueban un ramo y luego al segundo semestre tienen una menor carga y aprueban regularmente los ramos reprobados con anterioridad. Luego, un estudiante nuevamente se encuentra con nuevos ramos, los cuales reprueba, provocando el comportamiento de subidas y bajadas en el indicador PAS. En este gráfico sólo se muestran las cohortes 2001 a 2004 debido a que las otras cohortes aún no han cursado todos los semestres a analizar. Además, este gráfico es revelador desde el punto de vista del comportamiento de este indicador, el cual está determinado por la cantidad de asignaturas aprobadas y la cantidad de asignaturas cursadas en un semestre, por lo que se presume la posibilidad de predecir por lo menos estos datos del próximo semestre dado el historial de un alumno.

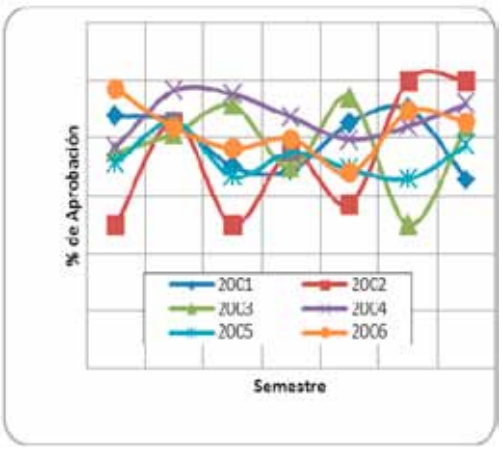

Figura 8. Porcentaje de Aprobación por Semestre (PAS) en asignaturas de la carrera (por semestre de permanencia para las cohortes de la 2001 a la 2004).

El gráfico de la Figura 9 muestra la tendencia de cantidad de asignaturas acumuladas por nivel sólo para alumnos de pregrado (que aún cursan ingeniería), donde se aprecia que la cantidad de asignaturas por nivel marca una tendencia muy parecida a los eliminados en cuanto a cantidad de asignaturas que adelantan, pero las asignaturas por segunda oportunidad son menores en comparación. En un contexto distinto, el gráfico de la Figura 10 muestra la tendencia de los promedios obtenidos tanto en PAA/PSU por año, lo que muestra que los puntajes en verbal e historia están por debajo del promedio general de ingreso a las carreras.

Resultados de la predicción mediante RNA

Considerando la arquitectura descrita en la Figura 4, los resultados de la estimación de la cantidad de asignaturas inscritas por un alumno se muestran en el gráfico de la Figura 11, para lo cual se realizó la estimación de los 563 datos de prueba, estimando sólo un semestre hacia el futuro. Para mayor claridad, la Figura 12 muestra un extracto de la misma Figura 11, donde se observa que la línea roja se sobrepone sobre la línea azul que representa la salida deseada, es decir, la salida que se pudo obtener desde el DW. Esto comprueba

experimentalmente que la predicción se ajusta bien a lo que la tendencia histórica del DW ha entregado, por lo cual el complemento entre DW y RNA es una herramienta potente para poder predecir el comportamiento futuro de un indicador de gestión.

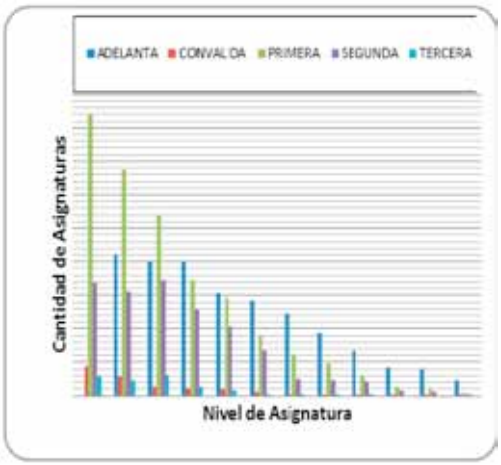

Figura 9. Gráfico Cantidad Acumulada de asignaturas que se cursan por nivel de asignaturas (solo pregrado).

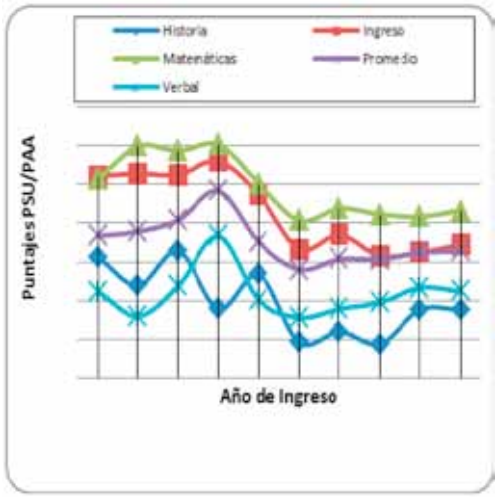

Figura 10. Gráfico Puntajes promedios de pruebas por año (todos).

En la Figura 13 se aprecia el gráfico que muestra la cantidad de asignaturas aprobadas por un alumno en un semestre determinado. Para una mejor apreciación en la Figura 14 se presenta un extracto de este gráfico donde se puede apreciar que la línea roja, que representa la salida de la predicción con RNA, se ajusta bastante bien a la salida deseada.

Como se puede observar en los gráficos de las Figuras 11, 12, 13 y 14, la información histórica obtenida del DW (línea azul) es muy similar a los valores predecidos por la RNA (línea roja). Con respecto a lo anterior, en la Tabla 1 se muestran los valores de los índices obtenidos para las estimaciones de ambas variables.

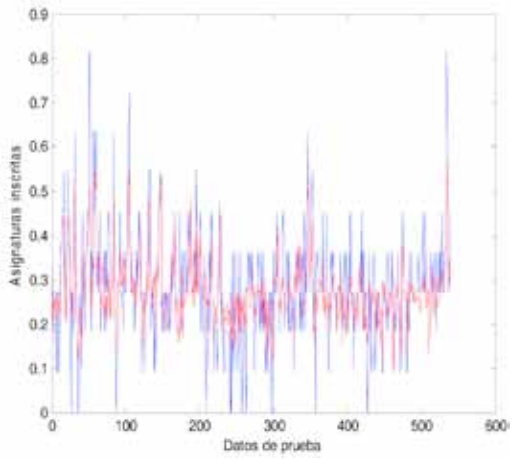

Figura 11. Salida estimada de las asignaturas inscritas en el siguiente semestre. La línea azul representa la salida deseada y la línea roja la salida estimada por la red neuronal. Los datos están normalizados entre 0 y 1.

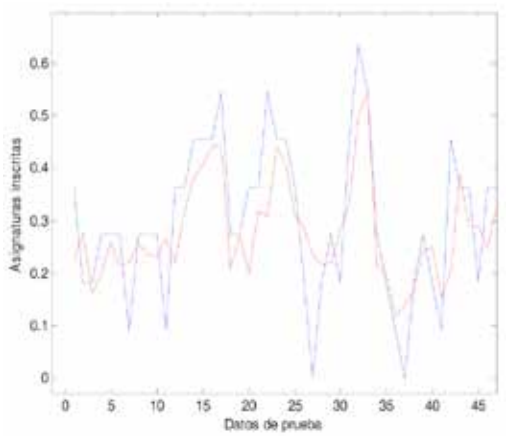

Figura 12. Extracto de la salida estimada de las asignaturas inscritas en el siguiente semestre. La línea azul representa la salida deseada y la línea roja la salida estimada por la red neuronal. Los datos están normalizados entre 0 y 1.

Tabla 1. Índices de adecuación y errores en la estimación de los datos de prueba.

Índices	Estimación de cantidad de asignaturas inscritas	Estimación de cantidad de asignaturas aprobadas
IA	0,7623	0,7180
RMS	0,2934	0,3678
RSD	0,0905	0,1225

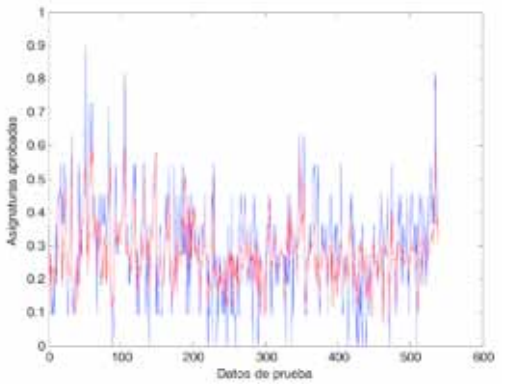

Figura 13. Salida estimada de la cantidad de asignaturas aprobadas en el siguiente semestre. La línea azul representa la salida deseada y la línea roja la salida estimada por la red neuronal. Los datos están normalizados entre 0 y 1.

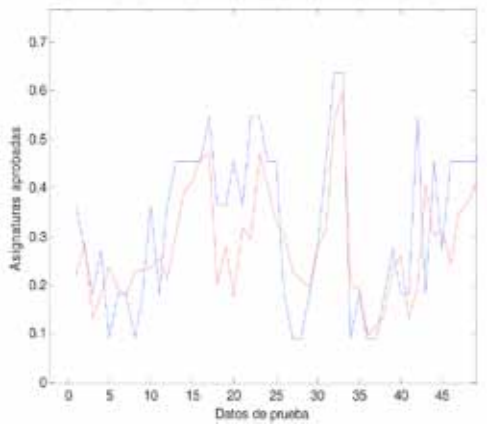

Figura 14. Extracto de la salida estimada de la cantidad de asignaturas aprobadas en el siguiente semestre. La línea azul representa la salida deseada y la línea roja la salida estimada por la red neuronal. Los datos están normalizados entre 0 y 1.

CONCLUSIONES Y TRABAJO FUTURO

Se ha realizado la implementación de un Data Warehouse y la implementación de una arquitectura de Red Neuronal Artificial para el análisis y la predicción de rendimiento académico de los estudiantes de Ingeniería Civil en Computación e Informática de la Universidad de Atacama. La principal ventaja en la utilización de un DW radica en la posibilidad de cruzar distintas dimensiones de análisis de forma simple y rápida, con tal de realizar un análisis exploratorio de los datos para la creación de reportes. Se puede destacar que el proceso de extracción, transformación y carga (ETL) es el que más tiempo y recursos demandó, debido principalmente a que la información debe ser cruzada desde distintas fuentes. Además, los sistemas operacionales no están diseñados para analizar datos y la heterogeneidad de las plataformas donde se encuentra la información añade una mayor dificultad que obliga a la creación de aplicaciones y sistemas específicos que permitan aprovechar los datos históricos. Es preciso agregar que la utilización de un modelo conceptual multidimensional para generar el esquema conceptual del DW se convierte en una gran herramienta que, independiente de las plataformas, permite acotar el dominio de análisis y dar claridad al proceso posterior de ETL.

Para finalizar respecto a la implementación de DW podemos indicar que el análisis mediante Rolap es eficiente y permite realizar operaciones en el cubo en tiempo real para poder navegar por los datos desde distintas perspectivas de una manera sencilla e intuitiva.

Por otro lado, se demostró cómo la arquitectura de RNA propuesta permite predecir el comportamiento del semestre posterior de un alumno respecto al semestre anterior en cualquier momento de permanencia en la carrera. A pesar de que los resultados de la aplicación de RNA pueden ser perfectibles utilizando información adicional del alumno, como información socioeconómica y encuestas, se piensa que la estimación alcanzada cumple con el objetivo de mostrar la tendencia futura en el comportamiento de un alumno.

En un contexto más genérico, es dable indicar que al obtener resúmenes y reportes usando DW, producto del análisis histórico de los datos, se puede crear

una base sólida de información para la arquitectura RNA y la predicción de comportamiento futuro. Con lo anterior, la utilización de DW más la utilización de técnicas de estimación o predicción (en nuestro caso una RNA) permiten un complemento para fundamentar análisis más completos pues como se muestra en este trabajo es posible predecir los indicadores de gestión que se obtienen del DW. Esto permite a la institución tomar medidas para poder analizar, modificar y validar los indicadores de gestión o quizás para generar nuevas estrategias que le permitan mejorar y/o optimizar su proceso de gestión, pues el conocimiento se extrae de sus mismas bases de datos, dando valor a la información de gestión que se registra pero que quizás no siempre se tiene en cuenta.

En el contexto particular de la Universidad de Atacama (UDA) se está en estos momentos en etapa de puesta en marcha de un nuevo sistema de gestión curricular, que tiene asociado el uso de la herramienta Cognus para la implementación de parte del enfoque presentado en esta investigación en toda la universidad una vez que se implante el sistema de gestión curricular. Es importante indicar que lo que se realizó en esta investigación fue para comprobar la utilidad de las herramientas y que tras comprobarse, la UDA decidió utilizar Cognus como solución BI. Otra decisión asociada a esta investigación tiene relación con acciones tomadas con el objetivo de mejorar el rendimiento académico, para lo cual se realizó una investigación sobre el uso y valoración de estrategias de aprendizaje [28] que complementó los resultados expuestos en este artículo; también se aplicó una investigación en aula sobre estrategia de codificación de información [26]. Con la combinación de los resultados [24, 25, 26, 27, 28] se decidió a nivel de diseño curricular en la carrera de Ingeniería Civil en Computación e Informática incluir la asignatura de estrategias de aprendizaje en la nueva malla curricular y generar un plan que permita la implantación de metodologías activas centradas en el alumno, todo lo anterior con el objetivo de mejorar el aprendizaje y rendimiento de los estudiantes.

Una posible debilidad del modelo podría darse si hay un cambio significativo en el contexto las predicciones que puedan dejar de ser válidas, al existir cambios en las políticas académicas, crisis social o económica, etc.

Como trabajos futuros se trabaja para generar indicadores con dimensiones sociales, económicas y dimensiones con datos de encuestas de perfiles biopsicosociales de los estudiantes con los que ya se cuenta, para generar una nueva arquitectura de DW. Además de la formalización de un proceso de inteligencia de negocios educacional.

AGRADECIMIENTOS

Este trabajo fue parcialmente financiado por la Dirección de Investigación de la Universidad de Atacama, Chile, Proyecto 221219 "Data Warehouse para Análisis con Jerarquías Difusas" de Carolina Zambrano Matamala. La autora también desea agradecer a la Sra. Marcela Varas Contreras por las sugerencias en el ámbito de Data Warehouse y al Sr. Gonzalo Acuña Leiva por las sugerencias en Redes Neuronales. Gonzalo Acuña Leiva desea agradecer al Proyecto Fondecyt 1090316 "Comparative Study of Support Vector Machines and Neural Networks for Nonlinear System Identification and Observer Design".

REFERENCIAS

[1] A. Bonifati, F. Cattaneo, S. Ceri, A. Fuggetta and S. Paraboschi. "Designing Data Marts for Data Warehouses". ACM Transactions on Software Engineering and Methodology. Vol. 10, Issue 4, pp. 452-483. October, 2001.

[2] L. Cabibbo and R. Torlone. "A Logical Approach to Multidimensional Databases". Lecture Notes in Computer Science. Vol. 1377. 1998.

[3] F. Carpani. "CMDM: Un Modelo Conceptual para la Especificación de Bases de Datos Multidimensionales". Tesis para optar al grado de Maestría. Universidad de la República. Uruguay. 2000. URL: http://www.fing.edu.uy/inco/pedeciba/bibliote/tesis/tesis-carpani.pdf.

[4] Z. Cataldi, F. Salgueiro y F. Lage. "Predicción del rendimiento de los estudiantes y diagnóstico usando redes neuronales". XIII Jornadas de Enseñanza Universitaria de la Informática. España. 2006.

[5] S. Chaudhuri and U. Dayal. "An Overview of Data Warehousing and OLAP Technology". SIGMOD Record. Vol 26, Issue 1, pp. 65-74. 1997. Pearson. 2004. ISBN 8420540250.

[6] E. Diaconescu. "The use of NARX neural networks to predict chaotic time series". WSEAS Transactions on Computer Research. Vol. 3, pp. 182-191. 2008.

[7] Y. Gao and M. Joo Er. "NARMAX time series model and prediction: feedforward and recurrent fuzzy neural network approaches". Fuzzy Sets and Systems. Vol. 150, pp. 331-350. 2005.

[8] M. Golfarelli, D. Maio and S. Rizzi. "Conceptual Design of Data Warehouses from E/R Schemes". Proceedings of the Thirty-First Hawaii International Conference on System Sciences. 1998.

[9] S. Haykin. "Neural Networks a Comprehensive Foundation". Second Edition. Macmillan College Publishing, Inc. USA. 1999. ISBN 9780023527616.

[10] B. Hüsemann, J. Lechtenbörger and G. Vossen. "Conceptual Data Warehouse Design". DMDW'00. Sweden. 2000.

[11] P. Isasi y I. Galván. "Redes de Neuronas Artificiales. Un enfoque Práctico". Pearson. 2004. ISBN 8420540250.

[12] C. Jiang and F. Song. "Forecasting chaotic time series of exchange rate based on non-linear autoregressive model". 2nd International Conference on Advanced Computer Control (ICACC). Shangai, China. 2010.

[13] R. Kimball, M. Ross and R. Merz. "The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling". John Wiley & Sons. 2002. ISBN 0471200247.

[14] U. Maulik and S. Bandyopadhyay. "Performance evaluation of some clustering algorithms and validity indices". IEEE transaction on patter analysis and machine intelligence. Vol. 24, pp. 1650-1655. 2002.

[15] J.N. Mazón, J. Trujillo, M. Serrano and M. Piattini. "Designing Data Warehouses: From Business Requirement Analysis to Multidimensional Modeling". In Proceedings of the 1st Int. Workshop on Requirements Engineering for Business Need and IT Alignment. Paris, France. September, 2005.

[16] T. Mitchell. "Machine Learning". McGraw-Hill. USA. 1997. ISBN 0070428077.

[17] G. Olguín. "Sistema de Monitoreo y Análisis del Comportamiento Académico del Alumnado". XXIII Congreso Chileno de Educación en Ingeniería. Concepción, Chile. 2009.

[18] Pentaho. "Pentaho Business Intelligence". URL: http://www.pentaho.com

[19] M.A. Pinninghoff, P. Salcedo and R. Contreras. "Neural Networks to Predict Schooling Failure/Sucess". Lecture Notes Computer Science. Vol. 4528. 2007.

[20] M.A. Pinninghoff, M. Herrera, R. Contreras and P. Salcedo. "Predicción de rendimiento académico mediante redes neuronales". VI Congreso Chileno de Educación Superior en Computación. Jornadas Chilenas de Computación. Arica, Chile. 2004.

[21] G. Salvendy. "Decision Support Systems". Handbook of Industrial Engineering: Technology and Operations Management. John Wiley & Sons, Chapter 4. 2001. ISBN 0471330574.

[22] C. Sapia, M. Blaschka, G. Höfling and B. Dinter. "Extending the E/R Model for the Multidimensional Paradigm". DWDM'98. Singapur, pp. 105-116. 1998.

[23] C. Todman. "Designing a Data Warehouse: Supporting Customer Relationship Management". Prentice Hall. 360 p. 2001. ISBN 9780130897121.

[24] C. Zambrano y D. Rojas. "Data Warehouse para analizar el comportamiento académico". XXIV Congreso Chileno de Educación en Ingeniería. Valdivia, Chile. 2010.

[25] C. Zambrano, D. Rojas, K. Carvajal y G. Acuña. "Data Warehouse y Redes Neuronales para el Análisis de Rendimiento de Alumnos: caso de Estudio con Alumnos de Ingeniería Civil en Computación e Informática de la Universidad de Atacama". XII Congreso Chileno de Educación Superior en Computación. Jornadas Chilenas de Computación. Antofagasta, Chile. 2010.

[26] C. Zambrano. "Propuesta metodológica y aplicación de estrategia de codificación de información a un curso de Introducción a la Programación". Congreso Chileno de Educación en Ingeniería. Valdivia, Chile. 2010.

[27] C. Zambrano, D. Rojas y M. Varas. "Data Warehouse con geolocalización y clustering". Congreso Internacional de Informática Educativa. Santiago, Chile. 2011.

[28] C. Zambrano. "Diseño, aplicación y análisis del uso y valoración de estrategias de aprendizaje y su relación con el rendimiento: Caso de estudio Ingeniería Civil Informática de la Universidad de Atacama". Congreso Chileno de Educación en Ingeniería. Aceptada 2011.