PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA
INGENIERÍA CIVIL INFORMÁTICA

PRONÓSTICO DEL DESEMPEÑO ESTUDIANTIL EN LA ESCUELA DE INGENIERÍA INFORMÁTICA DE LA PUCV

RICARDO ALFONSO BOCAZ LEÓN

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN INFORMÁTICA
PROFESOR GUÍA: RODRIGO ALFARO
PROFESOR CORREFERENTE: CRISTIAN RUSU

DICIEMBRE, 2013

Resumen

La continua búsqueda de la excelencia académica por parte de todas las instituciones de educación, han llevado a éstas a formularse preguntas sobre qué acciones debiesen ser tomadas para mejorar la calidad de la educación. Si se tuviera información anticipada pudiese ser aún mejor, ya que permitiría la toma de decisiones en forma oportuna por parte de la dirección de estas instituciones. Es por esto que el desarrollo de una herramienta que permita el pronóstico del desempeño estudiantil cobra importancia en este contexto y daría un apoyo a la asignación de recursos por parte de los directivos.

En esta memoria se detallan las bases para construir la herramienta antes mencionada mediante la utilización de máquinas de aprendizaje. Para lograr ésto se realizó un proceso previo de análisis de datos apoyado por herramientas informáticas disminuyendo los tiempos y aumentando la precisión. Posteriormente se compararon diferentes modelos predictivos para la clasificación de la finalización de las asignaturas de los estudiantes. El modelo que dió mejor resultado fue un Perceptrón Multicapa, llegando a una efectividad del 63,60 % y una exactitud de 64,01 % . La complejidad inherente del problema y los factores que no se incluyen en este proyecto y que a veces no pueden ser medidos, no permiten una clasificación más precisa.

Palabras Clave: desempeño académico, análisis inteligente de datos, minería de datos.

Índice

Lista de Figuras

Lista de Tablas

1 Introducción

En las últimas décadas, la cantidad de estudiantes que siguen programas de educación superior ha ido en aumento[1]. Este incremento de la población universitaria va acompañado por una expansión en la diversidad de las características de los alumnos. Estos últimos al provenir de diferentes sectores socioeconómicos, regiones y realidades tienen también necesidades y potencial académico distintos[1]. El reto de las universidades es reconocer esta heterogeneidad y afrontarla de la mejor manera posible. Power, en el año 1987, señaló que «el enfoque no solo debe estar en la admisión de un amplio abanico de estudiantes, sino también en darles el apoyo y la ayuda necesaria para asegurarles una oportunidad de éxito razonable»[2].

El desempeño académico de los alumnos constituye un indicador que permite aproximarse a la realidad educativa de la universidad y evaluar la calidad de su enseñanza. El avance del conocimiento, la fluidez en la transmisión de la información y los cambios acelerados de las estructuras sociales han incidido en el progreso alcanzado por los estudios enfocados en el rendimiento académico en educación superior [3]. Por lo tanto aumentar el desempeño de los estudiantes es el objetivo principal de cualquier institución educativa.

En la búsqueda de la excelencia académica la universidad se ha preocupado por los procesos formativos de sus estudiantes. Éstos procesos pueden ser vistos en términos de valor añadido, según el cual la calidad de una institución se estimaría por la diferencia entre las características de entrada y de salida de los alumnos[4]. La tendencia habitual de medir el proceso de egreso de un estudiante es correlacionar rendimiento con resultados, los cuales pueden ser clasificados como inmediatos o diferidos. Los resultados diferidos miden la utilidad que proporcionan los estudios para la inserción del titulado al mundo laboral, existiendo dificultades en su medición debido a lo subjetivo de ésta. Mientras que los rendimientos inmediatos se cuantifican con mayor facilidad, aunque no existe un estándar para su medición, ya que se pueden medir en términos del éxito. El rendimiento inmediato, entonces, queda relacionado a la superación de las exigencias de las asignaturas, y éstas a su vez como exigencias para culminar el programa de estudios, es decir: aprobación de las asignaturas contempladas en la malla curricular.

Conocer los diferentes factores que inciden en el desempeño académico en el campo de la educación superior de una manera integral, permite obtener resultados tanto cualitativos como cuantitativos para propiciar un enfoque completo en la toma decisiones, y así mejorar los niveles de pertinencia, equidad y calidad educativa[3]. Si se determinan factores comunes en grupos de estudiantes, estaríamos en presencia de patrones que describirían a los estudiantes y permitirían predecir su desempeño.

La Pontifica Universidad Católica de Valparaíso (PUCV), al estar inserta en el Consejo de Rectores de Universidades Chilenas (CRUCH), selecciona a sus postulantes solamente mediante la batería de la Prueba de Selección Universitaria (PSU). Contreras (2009) ha cuestionado esta metodología, determinando que el puntaje obtenido en la PSU no es el único indicador del

[1]Según estudio del Sistema de Información para la Educación Superior (SIES) http://www.mineduc.cl/usuarios/1234/File/Publicaciones/Estudios/5Estudio-Evolucion-Matricula-Historica-1990-2009.pdf

rendimiento posterior[5]. Para la Escuela de Ingeniería Informática de la PUCV sería una ventaja conocer si existen factores que influyen en el desempeño estudiantil aparte del puntaje PSU, así podría tomar decisiones eficientes con respecto a los recursos y sus asignaciones para mejorar el proceso formativo de sus estudiantes.

La PUCV dispone de un registro académico de las asignaturas dictadas en las carreras de la Escuela de Ingeniería Informática, así como de los estudiantes inscritos en ellas. Esta base de datos con información histórica es propicia para realizar un proceso de análisis de datos. Este proceso se apoyaría en las nuevas tecnologías para obtener conclusiones de manera rápida y precisa , que permitan el desarrollo de máquinas de aprendizaje. Para los docentes de la Escuela sería de vital importancia contar con información acerca de que características de los estudiantes influyen en mayor medida a su rendimiento y éxito estudiantil, es por esto que la creación de un modelo predictivo cobra importancia en este contexto ya que proporcionaría información para adelantarse a los hechos y así mejorar la asignación de recursos y por lo tanto la calidad de sus procesos formativos.

En la presente memoria se describirá cada una de las etapas necesarias para el desarrollo del trabajo. En la sección 2 se presentan tanto el objetivo general como los objetivos específicos del trabajo. El marco teórico del proyecto se encuentra descrito en la sección 3, consta tanto de los conceptos involucrados (subsección 3.1) como de las investigaciones relacionadas (subsección 3.2). El desarrollo de la memoria se encuentra en la sección 4, para finalizar con las conclusiones descritas en la sección 5.

2 Objetivos y Plan de Trabajo

A continuación se identifican los objetivos del presente trabajo, su alcance y las actividades desarrolladas en el transcurso de éste.

2.1 Objetivo General

Pronosticar el desempeño académico de los estudiantes de la Escuela de Ingeniería Informática de la Pontificia Universidad Católica de Valparaíso.

2.2 Objetivos Específicos

- Realizar un estudio acabado de los conceptos y las investigaciones que se relacionan con el presente proyecto.
- Realizar un análisis inteligente de datos para la identificación de características de los estudiantes que influyen en su desempeño.
- Aplicar técnicas de minería de datos para pronosticar el desempeño estudiantil.

2.3 Alcance

El alcance de este trabajo es el pronóstico del desempeño académico de los primeros años de carrera de los estudiantes de Ingeniería Civil Informática entre los años 2004 y 2010. Este trabajo se limita a este alcance debido a la poca información que existe de los estudiantes de Ingeniería en Ejecución Informática y a que no existe data histórica de cursos antes del 2004.

2.4 Planificación

Al comienzo de este trabajo, se planificaron las actividades a desarrollar. A continuación un detalle de estas actividades.

Definición de los objetivos del proyecto

Esta tarea contempló la definición de los objetivos del presente estudio. Los objetivos fueron confeccionados por el estudiante con ayuda del profesor guía.

Estudio del marco conceptual

La presente tarea involucraba la investigación de cada uno de los conceptos involucrados en el marco teórico: la minería de datos, el análisis inteligente de datos y el desempeño académico.

Obtención de los datos necesarios

La recolección de los datos necesarios para desarrollar la presente investigación consistió en dos subtareas: la definición de los datos que son necesarios así como también la solicitud de los datos a la Universidad y su recepción.

Realización del proceso de Extracción, Transformación y Limpieza (ETL)

Las primeras fases del análisis inteligente de datos: la extracción, limpieza y carga de los datos para su posterior análisis.

Realización del análisis Relational Online Analytical Processing (ROLAP)

Consistió en confeccionar los cubos que se utilizarán para el análisis y en examinar los datos con los cubos elaborados.

Preprocesado de los datos

Consistió en preparar los datos para la minería.

Data Mining

Se construyeron modelos predictivos, se entrenaron y registraron sus resultados.

Medir el desempeño de los modelos

Consistió en rescatar métricas que permitieron comparar los modelos y qué tanto poder predictivo poseían.

3 Marco Teórico

En este capítulo se entregará el marco referente al estudio, los conceptos que enmarcan el tema y los últimos estudios realizados que se relacionan a la presente investigación. Para comenzar se describirá el marco conceptual, que hace referencia a los conceptos principales que se relacionan al presente estudio, estos son, el desempeño estudiantil, el análisis inteligente de datos y la minería de datos. En el apartado 3.2 se detalla el marco referencial, que consiste en los estudios realizados con relación al tema que se está investigando.

Dado el objetivo de la presente investigación, es necesario contextualizar las bases del presente proyecto. La predicción, y por consecuente, la confección del modelo predictivo requiere de una serie de pasos anteriores a éste. Tales pasos son, por ejemplo, la definición de los datos necesarios para generar el modelo, el análisis inteligente para obtener conclusiones y la selección y prueba de las herramientas predictivas.

La definición de los datos requiere que previamente se haya pensado en qué indicadores de desempeño se quieren observar. También es necesario que los datos requeridos hayan sido guardados previamente por la institución. Estos datos son luego extraídos por alguna herramienta automatizada, ya que se requiere que esta extracción sea periódica. La periodicidad de la obtención debería ser en este caso semestral. Luego de la extracción de los datos, estos deben ser limpiados y transformados a conveniencia del análisis. Seguidamente se crea un almacén de datos con la idea de explorar a través de dimensiones y sacar conclusiones.

Todas las consideraciones antes descritas están relacionadas entre sí, como se puede ver en la figura 3.1. Con la data histórica guardada por la PUCV se realiza un proceso de Extracción, Transformación y Carga (ETL); que permite crear un almacén de datos desde los cuales es posible realizar un análisis ROLAP. Este análisis se ayuda en los indicadores de desempeño que se desean medir. El análisis inteligente de datos permite realizar un modelo predictivo que servirá de apoyo para la toma de decisiones de la dirección de la Escuela de Ingeniería Informática de la PUCV. Estas decisiones permitirán aumentar el desempeño, tanto de los estudiantes que se encuentran en la universidad, como también de los que próximamente se integrarán.

Para poder apoyar la toma de decisiones se hará uso de las herramientas de la minería de datos, para extraer conocimiento desde la información recabada por el análisis inteligente de datos que se aplicará. Esto permitirá pronosticar el desempeño futuro de los estudiantes de la Escuela, como también predecir el desempeño del primer año de los nuevos estudiantes que postulen a las carreras que dicta la Escuela.

Todo esto se apoya en las investigaciones que se han realizado a la fecha que intentan explicar los factores que influyen en el rendimiento universitario. Solamente algunas de ellas son detalladas en este documento, a razón de resumen. Aunque no se hayan encontrado hasta ahora investigaciones que utilicen el análisis inteligente de datos para explicar el desempeño universitario, estas investigaciones presentan una motivación para explicar de otra forma y como no quizás encontrar nuevas formas de poder predecirlo.

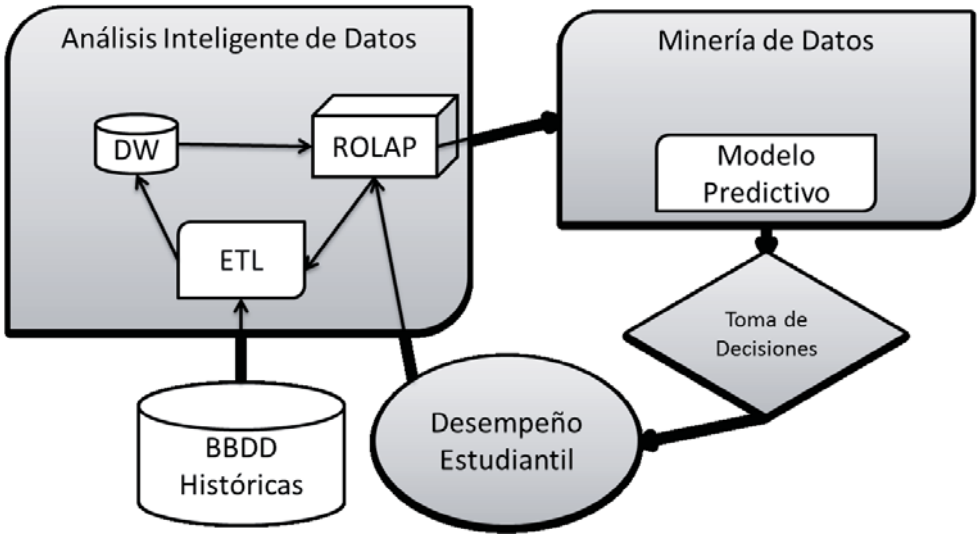

Figura 3.1: Relaciones entre los conceptos del proyecto

3.1 Marco Conceptual

En este apartado se pretende explicar a grandes rasgos los conceptos involucrados en el tema de investigación. Estos conceptos ayudan a comprender mejor la problemática a solucionar y permite focalizar la investigación. Se describe en un comienzo el desempeño estudiantil, seguido del análisis inteligente de datos, para culminar con la minería de datos.

3.1.1 Desempeño Académico

Para toda institución de educación es de importancia la medición del desempeño de sus estudiantes, ya que éstos son uno de los productos principales que crea la universidad. El continuo afán de mejorar los procesos de aprendizaje y la calidad de la educación, lleva a las universidades a preguntarse dónde se encuentran las falencias en su forma de realizar docencia. Para esto, una clara medición del rendimiento de sus estudiantes puede dar luces de donde se encuentran los principales factores, que hacen que un estudiante tenga un desempeño más elevado que otro.

El desempeño estudiantil se puede definir como el grado de cumplimiento de los objetivos planificados por el estudiante al iniciar el proceso educativo, comparado con los recursos disponibles para cumplirlos. También puede ser visto como la utilización eficiente de los recursos disponibles por el estudiante para lograr o sobrepasar los objetivos propuestos. Ejemplos de desempeño universitario es el tiempo que demora el estudiante en egresar, o la cantidad de créditos aprobados por sobre lo mínimo exigido.

Los factores que pueden influir en que un estudiante tenga un mejor desempeño son variados y afectan de diversas formas a cada estudiante. Existen factores que son propios de cada uno, como el conocimiento acabado de la carrera que estudia, o la motivación por el área de estudio que realiza. También factores externos, como los socioeconómicos, culturales, de entorno, entre otros. También se pueden apreciar componentes propios del ambiente universitario, tales como la infraestructura, sus profesores, el compromiso con el estudiante, las ayudas por parte de la institución, los compañeros de clase, etc. Al final, encontrar todos los factores que podrían

incidir es una tarea titánica, sino imposible, para cualquiera que desee realizarla.

El rendimiento académico en estricto rigor no es lo mismo que el desempeño académico. El primero mide la relación de los objetivos propuestos por el estudiante con lo que realmente cumple, mientras el segundo hace referencia a la buena utilización de recursos tales como el tiempo, dinero, esfuerzo, etc. en lograr los objetivos propuestos. Es por ésto que los dos conceptos no deben ser confundidos, ya que el rendimiento mide la eficacia de los objetivos del estudiante mientras que el desempeño mide la eficiencia de éstos. Aunque podría establecerse que el desempeño engloba al rendimiento estudiantil, ya que al medir la eficiencia de los recursos en cumplir el objetivo necesariamente los objetivos deben cumplirse. Es por esto último que este estudio está enfocado en el desempeño universitario, sin embargo, también se medirán aspectos de eficacia.

Para tener una idea de la eficiencia, hace falta definir como se medirá ésta. Esto se realiza a través de la construcción de indicadores de desempeño, que darán un diagnóstico de la situación del estudiante. La confección de estos indicadores pasa por un discernimiento por parte de la dirección de la institución sobre qué realmente desea medir para su posterior mejora. En la literatura se pueden encontrar diversos indicadores, que han sido construidos con alguna realidad particular, pero que no necesariamente se aplican a cualquier tipo de investigación.

3.1.1.1 Indicadores de desempeño

Básicamente los indicadores que se construyan deben dar una idea de la eficiencia de algún aspecto educativo, para esto se debe especificar primero qué se desea medir y cuales son las unidades que indican lo que se desea medir. Existen de distintos tipos, como las que miden las calificaciones o esfuerzo que un estudiante realiza para terminar su carrera, los que miden el tiempo que se emplea para aprobar una asignatura, un grupo de asignaturas o lograr el egreso, o también los que miden la aprobación con respecto a la carga académica de un año en particular.

Las calificaciones puede ser medidas de diversas formas, y dependerá del constructor del indicador velar que éste mida lo que realmente se busca. Un ejemplo sería medir el promedio de calificaciones por año, o ponderarlas por los créditos de cada asignatura, otro podrá ser medir el promedio de calificaciones de un grupo de asignaturas o el promedio ponderado acumulado, etcétera. Los indicadores que tienen relación con el tiempo pueden medir distintos aspectos de éste, por ejemplo el tiempo de egreso, el tiempo en que se demora un estudiante en aprobar un grupo de asignaturas o una asignatura en particular, el tiempo que se está sin estudiar (congelamiento), el tiempo que transcurre entre que el estudiante debe tomar una asignatura y realmente la inscribe, el tiempo que demora el estudiante entre que egresa y se titula, la diferencia de tiempos entre que el alumno egresa y la duración de la carrera, entre otros. Los indicadores que midan la aprobación pueden indicar, por ejemplo, el porcentaje de aprobación de un área de estudio, el porcentaje de aprobación de las asignaturas aprobadas con respecto a las inscritas, la relación entre los créditos aprobados y la tasa de avance, promedio de créditos aprobados por semestre, y un largo etcétera.

Por supuesto que pueden construirse otros indicadores que no sean de los tipos anteriormente descritos, siempre que vengan a responder necesidades por parte de la Dirección y que puedan aportar información que permita describir la situación actual de la Escuela. Estos indica-

dores se retroalimentan y pueden modificarse, añadiendo o quitando parámetros para ajustarlos a las necesidades que surjan. También se pueden dejar de calcular algunos indicadores que no aporten en un momento determinado a medir la eficiencia de los estudiantes en cuanto a su formación académica.

3.1.1.2 Rendimiento académico

No es menos importante medir la consecución de los objetivos planteados por los estudiantes, y como no por la misma Escuela, ya que permite ver en que aspectos los estudiantes no cumplen con sus objetivos o en que porcentaje lo hacen, etc. Los indicadores que podrían medir el rendimiento académico deben estar en concordancia con los objetivos que se plantean los constructores de ellos.

Como ejemplos de indicadores que pueden medir el rendimiento académico están la relación entre la cantidad de créditos aprobados con los del total de la carrera, el total de asignaturas aprobadas con relación al total, la tasa de avance, entre otras. También podrían ser medidos indicadores por parte del grupo de estudiantes, como por ejemplo los de una cohorte en especifico y calcular los indicadores descritos anteriormente. Con esto se podría medir el rendimiento de la cohorte produciendo información que no puede ser vista por el cálculo de los indicadores a sujetos individuales.

Es necesario recalcar que ni el rendimiento ni el desempeño académico son conceptos que tienen un criterio aceptado por todos. constituye un constructo que puede ser hecho operativo de distintas maneras en función del significado que tiene para cada sujeto de acuerdo con su situación particular. De ahí que se considere este término como un concepto multidimensional, relativo y contextual.[4]

Desde un punto de vista práctico la tendencia más habitual es identificar rendimiento con resultados, distinguiendo dos categorías: los inmediatos y los diferidos. Los rendimientos inmediatos hacen relación con la estadía de los estudiantes en la universidad, como se han tratado durante toda esta sección. En cambio los diferidos son los resultados que obtienen los estudiantes cuando ya dejan la universidad.

Los rendimientos diferidos presentan un grado de dificultad en su cálculo, debido a que no siempre se dispone de la información de la vida posterior a la universitaria de cada estudiante, además de que los indicadores que podrían usarse no presentan la facilidad de la elaboración que tienen los del rendimiento inmediato. El rendimiento diferido hace relación con el impacto que tiene la enseñanza universitaria sobre la vida social. Por lo tanto puede medir distintas variables como las socioeconómicas -capacidad del estudiante en insertarse al mundo laboral, nivel de salario, estatus, retribución- como también variables del tipo personal y social como por ejemplo el desempeñar trabajos que tengan que ver con su título, estabilidad de los empleos, etc.[6]

3.1.2 Análisis Inteligente de Datos

El análisis inteligente de datos (AIDA), como se muestra en la figura 3.2, es una metodología que proviene de la complementación de diferentes disciplinas, tales como la Estadística y las Máquinas de Aprendizaje[7]. La Estadística aporta el modelado de la realidad, la recolec-

ción de los datos, el análisis de estos y el enfoque de sacar conclusiones a partir de una muestra de datos (inferencia). Para realizar lo anterior se apoya en las herramientas de dos ciencias: las Ciencias Sociales y las Ciencias Matemáticas. Mientras que las Máquinas de Aprendizaje se nutren, además de las Ciencias Matemáticas, de las Ciencias de la Computación. Las Máquinas de Aprendizaje permiten potenciar el análisis de datos generando computadores que aprenden y pueden calcular grandes cantidades de información en un menor tiempo comparado con lo que demoraría una persona.

El análisis inteligente de datos no es más que un análisis de datos apoyado en las nuevas tecnologías. Éstas permiten realizar el estudio con mayor rapidez y precisión, disminuyendo los errores. Además el AIDA permite procesar cantidades de datos de tamaño superior. Existen herramientas genéricas que permiten realizar un AIDA, y hasta se han creado metodologías o conjuntos de buenas prácticas basados en este tipo de análisis, como por ejemplo la Inteligencia de Negocios. Las aplicaciones del AIDA pueden desarrollarse en variados campos, tales como las Ciencias de la Ingeniería, la Economía, etc.

3.1.2.1 Análisis de Datos

Los primeros pensantes y científicos se interesaron por el tratamiento matemático de los datos y la información. Desde Galileo, pasando por Platón y los analistas de datos y teóricos tales como Jean-Paul Benzécri. Él y Gödel, eran conscientes de las dificultades de la observación o la medición de los datos, y los factores explicativos subyacentes a los fenómenos asociados con los datos. Con la llegada de los computadores en la actualidad es posible realizar la recolección de grandes escalas de datos, y de analizar y modificar cantidades de información de gran tamaño, lo que ha dado origen al análisis de datos moderno.[8]

El análisis de datos se define como «El proceso de computar diversos resúmenes y valores derivados de la colección de datos dada»[7]. El termino *proceso* dice relación con lo que es en sí el análisis de datos, algo iterativo, incremental y con *feedback*; con diferentes entradas y salidas. Es iterativo porque se examinan los datos, se aplica una técnica analítica y se decide probar de otra forma, quizás modificando los datos, dividiéndolos, o en algunos casos transformándolos; y se vuelve al comienzo aplicando otra técnica analítica. Ésto puede ser realizado repetidas veces durante el análisis, por que cada técnica puede probar aspectos distintos de los datos, aspectos que con una sola técnica no podrían haber sido abarcados y que sirven para responder otras preguntas que se generen en el camino. Las herramientas que se utilizan en el análisis de datos no deben ser consideradas como técnicas aisladas de las demás, sino todo lo contrario, porque tienen complejas relaciones entre ellas.

Se pueden distinguir dos tipos de análisis, los descriptivos y los inferenciales. Los primeros apuntan a rescatar información propia de la muestra que se analiza ,y responder preguntas como por ejemplo ¿Cuál es la edad promedio de los asistentes de este partido de fútbol?, ¿Cuántos bebés nacieron con el pelo rubio ayer en Valparaíso?, etc. En cambio los segundos tratan de explicar como se comporta la población basado en una muestra representativa de ella, por lo tanto responden preguntas amplias, como por ejemplo ¿Cuántos niños rubios nacieron en el mundo ayer?, que se infieren porque no se puede analizar a toda la población.

Las herramientas utilizadas en el análisis de datos difieren si es que se habla de un análisis

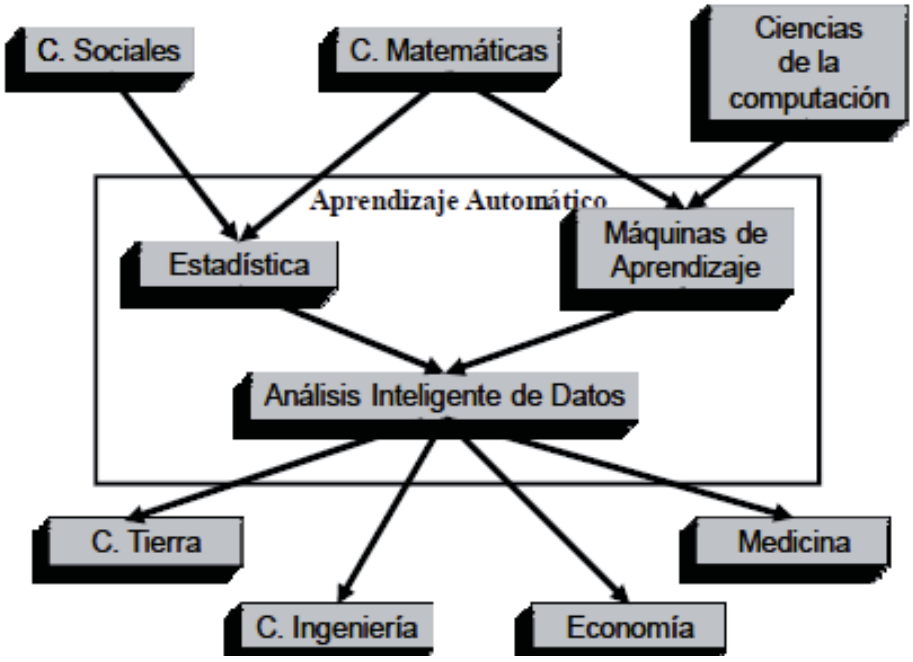

Figura 3.2: Ciencias que aportan al análisis inteligente de datos

descriptivo o de uno inferencial. Aunque puede ser la misma herramienta que se utilice para el análisis, la comprensión del resultado es distinto. Si se toma, como un ejemplo, la media de una muestra y se realiza un análisis descriptivo, entonces es un valor exacto y calculable; en cambio si se infiere la media, siempre será un valor estimado y si la muestra es modificada, la media estimada también cambiará, por lo tanto siempre se está consciente *a priori* de que existe un error en la estimación.

La nueva forma de hacer estadística, asistida por computador, ha llevado a que el análisis de datos se preocupe de descubrir conocimiento en forma confiable, rápida y oportuna. Ya no se necesita de tanto tiempo realizando tablas de datos a papel y lápiz; calcular parámetros estadísticos y descubrir conocimiento es algo que se realiza con rapidez y que permite procesar una gran cantidad de datos, que sin la ayuda del computador no podría ser posible. Los modelos de aprendizaje automático permiten realizar una minería de datos precisa y óptima, aportando al descubrimiento de conocimiento un mayor potencial del que ya tenia.

3.1.2.2 Los datos y su naturaleza

Para realizar un análisis de datos lo primero es determinar la naturaleza de los datos. La mayoría de las veces los datos son del tipo numérico, pero existen otros tipos de datos como por ejemplo las imágenes y los textos. Es necesario saber de qué tipo son los datos, para así poder elegir las técnicas apropiadas para su análisis. Una identificación de los tipos de datos permitirá una mejor representación de ellos. Existen tres categorías de datos, las cuales se diferencian por su estructura y representación: los estructurados, los no estructurados, y los bloques de datos binarios.

Los datos estructurados pueden ser subclasificados también en lo que son los datos cualita-

tivos, los cuantitativos, los simbólicos y los que están ordenados jerárquicamente. El tipo de dato estructurado está dado por su significado para las personas, los datos cualitativos denotan atributos característicos de los objetos que representan sus cualidades, mientras que los cuantitativos representan mediciones de esos atributos de forma objetiva y exacta. Los datos estructurados son representados por estructuras de datos tales como los arreglos, las tuplas, las matrices, árboles, etc.

Los datos no estructurados son principalmente textos. Un texto está compuesto por unidades fundamentales que son las palabras y de símbolos que las acompañan. Estas unidades pueden ser combinadas de maneras más diversas que los números. Por último, los bloques de datos binarios, hacen relación a la representación en forma computacional o electrónica de imágenes o sonidos.

3.1.2.3 Aplicaciones del análisis de datos

El análisis de datos da soluciones a problemas propios de clasificación, regresión, reconocimiento y transformación. Las disciplinas que se ven favorecidas con su aplicación son diversas. Algunas de ellas son la biometría, climatología, sismología, medicina, balística entre otras.

Las aplicaciones más comunes son:

- Pronóstico de magnitudes máximas de terremotos.
- Pronóstico de perspectiva de yacimientos minerales.
- Pronóstico de tormentas en la ionósfera.
- Regionalización sísmica.
- Diagnóstico diferencial de enfermedades.
- Lectura diagnóstica de señales (EEG, ECG,IC, etc.).
- Clasificación automática de clientes.
- Identificación de huellas digitales.
- Identificación de caligrafías.
- Identificación de rostros (estáticos, en movimiento, enmascarados, etc.).
- Identificación de objetos(mediante sonidos o mediante rastros).
- Dispositivos de acceso por identificación iriológica.
- Reconocimiento de placas de vehículos.
- Caracterización sociopolítica de colectivos sociales.
- Pronóstico de surgimiento de fenómenos sociales.

- Caracterización del *modus operandi* de terroristas o delincuentes.
- Análisis de las causas de algún fenómeno social.

3.1.2.4 Etapas del Análisis Inteligente de Datos

En el Análisis Inteligente de Datos existen dos etapas. La primera es la integración de los datos, mientras que la segunda el análisis propiamente tal. La integración se convierte entonces en una actividad de soporte, en la cual el objetivo principal es llenar los almacenes de datos para posterior estudio. La integración de los datos tiene, a su vez, etapas definidas en cuanto al tipo de actividad que se realiza con los datos. Estas etapas, como puede verse en la figura 3.3 son la extracción, la transformación y la carga de los datos (ETL).

El proceso de ETL puede ser ejecutado de dos maneras distintas: como transformaciones y como trabajos (jobs). Las transformaciones se realizan en demanda, esto es, cuando se requiere realizar una ETL se ejecutan los pasos cíclicamente. En cambio los trabajos o *jobs* son procesos automáticos que se activan en un determinado momento o cuando se logra una condición determinada del origen.

3.1.2.5 Integración de los datos

La denominada integración de datos consiste en el proceso de llenado de los almacenes de datos. Para realizar este proceso se utilizan herramientas automatizadas. Este proceso, aunque está dividido en fases no debe considerarse como pasos secuenciales, ni tampoco como requisitos para el llenado de los almacenes. La integración de datos consiste en la extracción, la transformación y la carga. Estas fases son iterativas e incrementales, existiendo actividades que apoyan estas fases.

Extracción: Es el proceso de obtener los datos desde el medio o fuente original, éstos pueden adquiridos desde un sensor si no se tienen ya registrados, o realizar alguna consulta a una base de datos o archivo donde estén contenidos. Un ejemplo de extracción es la obtención del registro de todas las ventas en un supermercado.

Transformación: Consiste en cambiar la forma o representación de los datos para hacerlos coincidir con la forma del almacén de datos que se desea crear. Por ejemplo eliminar el dígito verificador a los registros del RUT de los clientes del supermercado.

Carga: Esta etapa radica en escribir los datos en el almacén creado.

Las actividades de apoyo al proceso de extracción se describen a continuación:

Captura de datos modificados: Consiste en solamente extraer los datos que hayan cambiado o se hayan agregado a la fuente después de la última extracción realizada.

Puesta en escena: A veces no es eficiente transformar directamente los datos extraídos, por lo que se opta por el almacenamiento temporal para su posterior transformación.

Para la etapa de transformación también existen actividades que apoyan esta labor, algunas de ellas son:

Validación: Es la verificación de que los datos de origen sean correctos y consistentes. Los datos que no lo sean pueden ser filtrados.

Limpieza: Consiste en el proceso de modificar los datos inválidos.

Decodificación: A veces los datos de origen no son aptos para mostrarlos en reportes, por lo tanto se deben modificar haciéndolos entendibles para el usuario final.

Agregación: Para aumentar el rendimiento en consultas que deban procesar gran cantidad de datos se prefiere agregarlos, esto es realizar operaciones para un conjunto de datos tratando de explicar el conjunto en su totalidad. Por ejemplo calcular promedios o sumas de elementos comunes. Esta actividad puede realizarse en la etapa de transformación, para que el análisis tenga un rendimiento superior.

Identificación: Consiste en generar y administrar nuevas claves primarias e indices para el almacén de datos. Al transformarse los datos y llevarlos a dimensiones o crear tablas de hechos, deben crearse a veces nuevas claves identificadoras.

En la etapa de carga se encuentran las siguientes actividades que apoyan el proceso:

Carga de tablas de hechos: Consiste en crear y llenar las tablas de hechos o *fact tables*. Se pueden crear nuevas filas en el almacén de datos o también modificar las existentes.

Carga de tablas de dimensión: Al crear nuevas filas de hechos puede también ser necesaria la creación o modificación de las tablas de dimensiones.

3.1.2.6 Procesamiento Analítico En Línea (OLAP)

Existen variadas formas de realizar un análisis de los datos. Específicamente en este proyecto se utilizará el análisis en línea relacional (ROLAP). La literatura nos presenta tres tipos de análisis OLAP, estos son el análisis multidimensional MOLAP, el relacional ROLAP y el híbrido entre las dos anteriores HOLAP.

Análisis MOLAP: Las tablas del almacén de datos son multidimensionales. La aplicación MOLAP consta de dos partes: las tablas multidimensionales y el motor de procesamiento. Su ideal es la eficiencia de conservar la información desde antes en dimensiones. Su característica principal es el alto desempeño.

Análisis ROLAP: Dado el uso masivo de las bases de datos relacionales, la corriente de pensamiento ROLAP cree que las dimensiones pueden ser guardadas como relaciones entre ellas. De aquí nacen los esquemas de almacén de datos tales como el estrella o el copo de nieve. Es el análisis de moda en la actualidad por su facilidad de uso y la proliferación de ambientes de código abierto. Se caracteriza por tener alta escalabilidad.

Análisis HOLAP: De desarrollo mas reciente, combina las dos visiones anteriormente descritas. Obtiene por lo tanto las ventajas de ellas, gran escalabilidad y desempeño.

Figura 3.3: Etapas del AIDA

Cubos, Esquemas, Dimensiones y Medidas El análisis ROLAP se basa en unidades fundamentales llamadas dimensiones. Éstas representan un elemento abstracto que modela la realidad a analizar. Dependiendo del problema pueden resultar diferentes dimensiones o ser agregadas o modificadas en el camino. Básicamente son tablas en una base de datos que agrupan características que el desarrollador del análisis considera que pertenecen a una entidad en particular.

Las dimensiones a veces pueden relacionarse con otras dimensiones. Para estos casos se debe definir qué esquema seguirá el almacén de datos. Existen esquemas de amplia utilización tales como el estrella o el copo de nieve, mostrados en la figura 3.4.

En las tablas de hechos, existen métricas cuantitativas o cualitativas. Estas descripciones son llamadas medidas. Las medidas son generalmente un dato del tipo numérico. Estos datos son los que se agrupan o agregan cuando se analizan los datos mediante ROLAP.

Todas estas entidades vistas anteriormente se integran entre sí formando lo que se denomina un cubo, que puede poseer tres o más dimensiones. Las «caras» de este cubo son las dimensiones que se generaron en el almacén. Cada dimensión se desagrega en niveles. Estos niveles se intersectan con los niveles de las otras dimensiones a través de las medidas. El cubo puede estar formado solamente por dimensiones anidadas, por lo que en el esquema de estrella cada cara corresponde a una dimensión.

3.1.3 Minería de Datos

La minería de datos despierta gran interés, ya que al tener como objetivo la obtención de conocimiento desconocido o predecir el futuro, parece que a muchos le podría parecer beneficioso. El término en si es utilizado indistintamente si se trata de máquinas de aprendizaje o de minería de datos en sí, aunque las máquinas de aprendizaje como concepto abarque un espectro mayor. Como definición de minería de datos se puede exponer que es «el proceso no trivial

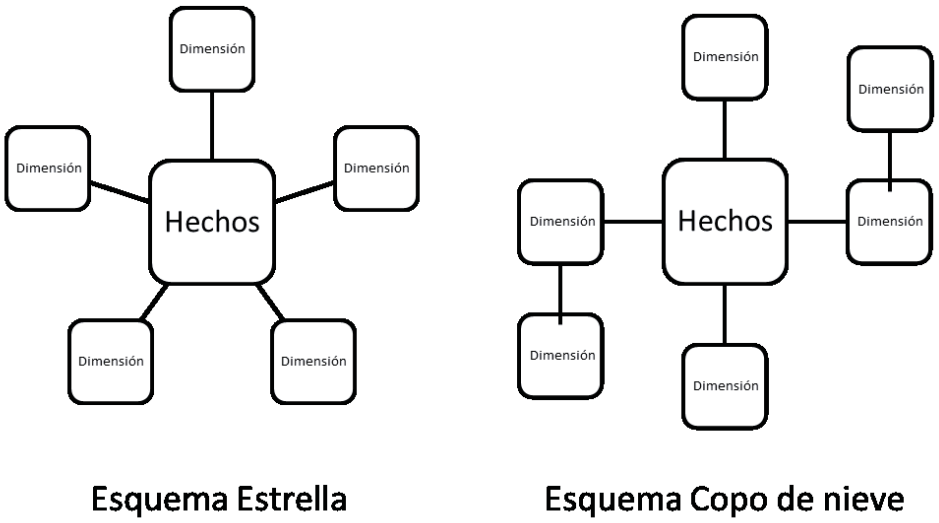

Figura 3.4: Esquemas de almacén de datos

de identificar patrones válidos, novedosos, potencialmente útiles y que en ultima instancia sean entendibles»[9]

Por lo tanto, basándose en la definición anterior podemos decir que el proceso de minería de datos debe tener una serie de características. Primero, los patrones que se deben identificar deben ser entendibles, por lo tanto los resultados que se obtengan deben tener sentido. Deben ser novedosos, esto quiere decir que digan algo que no se sabía con anterioridad y que los resultados deben ser válidos, dado un cierto contexto.

3.1.3.1 El Proceso de Minería de Datos

La minería de datos es un proceso, y tiene mucho en común con el análisis inteligente de datos. Existe un proceso documentado de minería de datos llamado CRoss Industry Standard Process for Data Mining (CRISP-DM)[2] esquematizado en la figura 3.5. Al igual que en éste, los datos son la parte central del proceso. Las actividades de entender el negocio, entender los datos y prepararlos son similares a las necesarias para construir un almacén de datos.

Las diferencias comienzan a notarse en la etapa de modelado. El modelado consiste en considerar diferentes modelos y elegir el que mejor se comporta. El objetivo de la minería de datos no es explicar las posibles relaciones entre los datos, sino que está enfocada en encontrar soluciones prácticas para predecir ciertas salidas.

3.1.3.2 Herramientas de Minería de Datos

La naturaleza predictiva de la minería de datos es la que la diferencia del AIDA. El AIDA se preocupa de el análisis de los datos históricos, y en comparar objetivos con las medidas actuales. La minería de datos provee herramientas para predecir estas medidas, con algún grado de confianza. Y las herramientas que utiliza son los modelos y los algoritmos. Existen cuatro categorías que permiten realizar minería de datos: clasificación, asociación, *clustering* o agru-

[2]Una completa referencia a este estándar puede verse en http://www.crisp-dm.org

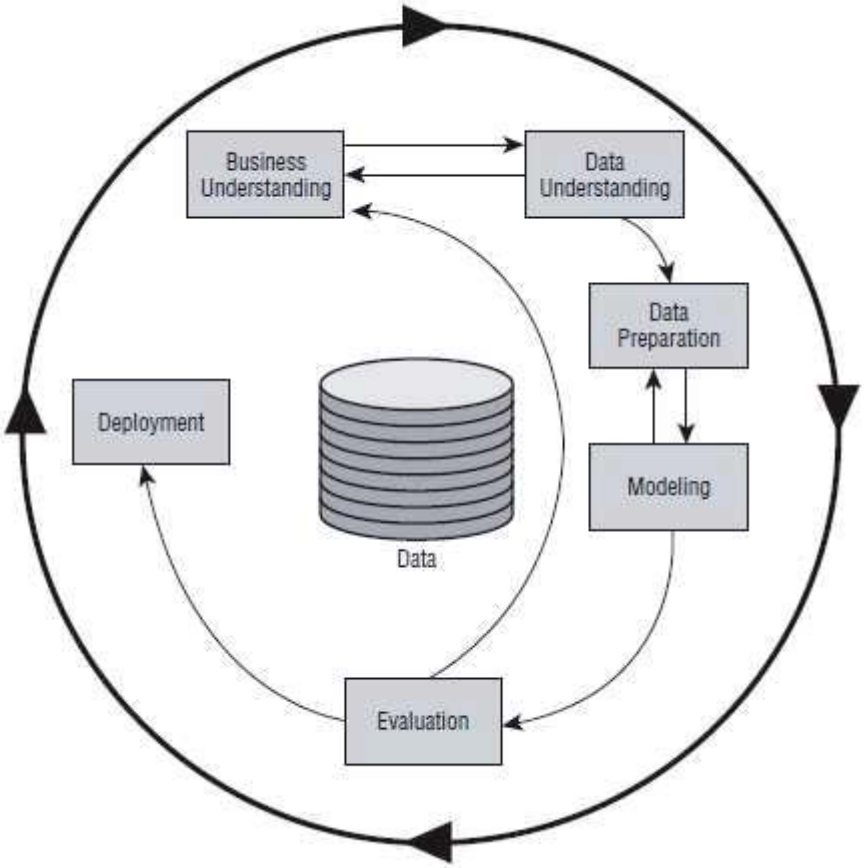

Figura 3.5: Esquema de CRoss Industry Standard Process for Data Mining (CRISP-DM)

pamiento y regresión[10].

Clasificación La clasificación es el proceso de separar o dividir el conjunto de datos en grupos llamados clases. Los miembros de cada clase deben ser homogéneos entre sí y heterogéneos con respecto a los miembros de las otras clases. La heterogeneidad y homogeneidad está definida en función de la distancia entre los elementos. Esta distancia es medida con respecto a la o las variables que se intentan predecir.

La clasificación entonces es asignar una etiqueta o clase a una entrada dado un conjunto de datos. Se suele tener un conjunto de datos que se les conoce su clase y se entrena a la máquina con estos datos. Un ejemplo de clasificación simple es la definición de a qué idioma pertenece cierta palabra, teniendo un conjunto de palabras de las cuales ya se sabe su idioma y poder predecir a qué idioma pertenece una nueva palabra no incluida en los datos de entrenamiento.

Asociación La asociación consiste en encontrar cuál es la relación entre dos o mas elementos de un conjunto de datos. Una aplicación de asociación comúnmente es la de los productos que se compran juntos en un supermercado; el famoso ejemplo de la cerveza y los pañales. La asociación explica correlaciones entre dos elementos pero no causalidad.

Clustering o Agrupamiento El *clustering* es similar a la clasificación. Se trata de encontrar que elementos de un conjunto de datos comparten características comunes. La diferencia está en que en el *clustering* no se sabe desde antes a qué clase pertenecen los datos de entrenamiento, en cambio en la clasificación si.

Esta diferencia también se denomina aprendizaje supervisado y no supervisado. Mientras que en la asociación el entrenamiento es del tipo supervisado, en el *clustering* es del tipo no supervisado y las clases las encuentra la misma herramienta. En el aprendizaje no supervisado la herramienta de minería de datos solamente encontrará grupos que comparten ciertas características y el investigador será el encargado de definir cuales son las razones detrás de eso.

Regresión La clasificación, *clustering* y asociación predicen clases especificas, que son valores nominales, es decir no numéricos. A menudo se requiere predecir una salida numérica basada en data histórica. Esto presenta una mayor complejidad, ya que las posibilidades son infinitas, tal como los números. A esto último se le denomina regresión. Existen diferentes formas de regresión tales como las lineales y las no lineales.

Entrenamiento y Pruebas Estos dos términos a menudo son confundidos. El entrenamiento es el proceso de construcción del modelo predictivo. Para que la máquina aprenda es necesario entrenarla con un conjunto aleatorio de datos y que tenga un tamaño considerable comparado con la totalidad del conjunto de datos. Las pruebas, en cambio, son las actividades

que verifican la validez y/o calidad del modelo construido. La data usada para estas dos actividades no debe ser la misma. El tamaño del conjunto de entrenamiento debería ser de dos tercios del total dejando el otro tercio para las pruebas.

3.2 Marco Referencial

Siempre que se comienza una investigación es importante revisar en la literatura si es que lo que se está investigando ya no ha sido realizado por alguien anteriormente, es por esto que se deben recopilar y estudiar las investigaciones que se relacionen con el tema para también poder basarse en lo ya estudiado por éstas y no reinventar la rueda. Es por esto que es necesario a continuación describir los estudios realizados sobre el tema que se han encontrado durante el transcurso de la investigación.

3.2.1 Determinantes de desempeño universitario. ¿Importa la habilidad relativa?

Este estudio fue realizado por Dante Contreras, perteneciente al Departamento de Economía de la Universidad de Chile; Sebastián Gallegos del Departamento de Estudios del Ministerio de Educación y Francisco Meneses procedente de la Universidad de Wisconsin-Madison y del Banco Central de Chile en el año 2009.

Esta investigación examinó si el haber tenido un buen desempeño relativo en el colegio de egreso de enseñanza media es un buen predictor de rendimiento universitario. La motivación de este trabajo fue revisar si una medida de habilidad relativa podía entregar información relevante y adicional respecto de las proyecciones académicas de los estudiantes, que no aportan los instrumentos de la batería de selección actualmente en uso.

Se examinó la relación entre el desempeño académico de los estudiantes de primer año en cuatro universidades - Pontificia Universidad Católica de Valparaíso, Universidad Católica de Temuco, Universidad Finis Terrae y Universidad de Santiago - y los instrumentos de selección tradicionalmente usados. Luego, se incluyó una medida de habilidad relativa: el *ranking* de egreso del alumno en su colegio, como un predictor de rendimiento adicional. Ello entregó evidencia respecto de cómo rendirían los alumnos con un determinado *ranking* en la universidad, condicional a haber ingresado bajo el sistema de selección tradicional. Este análisis permitió tener una evaluación ex-ante de una política de acceso a la educación superior que considere al *ranking* de egreso como un instrumento de selección.

También se analizó el rendimiento de alumnos ingresados mediante un sistema especial (o cupos supernumerarios) a la universidad. Estos cupos se entregaron a los alumnos que, perteneciendo al 5% de mejor rendimiento de su colegio de egreso, hubiesen quedado marginalmente por debajo del punto de corte de la carrera a la que postularon. Lo anterior entrega evidencia acerca del desempeño escolar relativo ex-post.

Los resultados indican que haber estado entre los mejores estudiantes de la escuela de egreso implica un mejor desempeño universitario en el primer año, aún controlando por los puntajes obtenidos en las PSU y las NEM para cada carrera. También se demostró que los estudiantes que ingresan por cupos supernumerarios obtienen rendimientos estadísticamente iguales

en el primer año cuando se controla por el puntaje de ingreso a cada carrera en dos de las tres universidades. Es decir, los alumnos que ingresan por el sistema especial siguen un patrón de rendimiento similar o levemente superior en el primer nivel universitario, que aquellos que ingresan por la vía tradicional. Los alumnos que ingresaron bajo los cupos supernumerarios, tuvieron un rendimiento igual o superior a lo que su puntaje de ingreso a la universidad hubiese predicho.

3.2.2 Búsqueda de patrones de rendimiento académico mediante técnicas de análisis multivariante. Aplicación a E4.

Esta investigación fue efectuada por Antonio Rúa de la Universidad Pontificia de Comillas en Madrid, en el año 2001. Se aplicó el estudio a los alumnos de 1º de E4 de la carrera de Ciencias Empresariales Internacionales impartida en la Facultad de Ciencias Económicas y Empresariales de la Universidad Pontificia Comillas de Madrid. El conjunto de datos se corresponde con las notas obtenidas en la convocatoria de junio del curso 1999-2000. Debe aclararse que para evitar la perdida de datos a aquellos alumnos que no se presentaron al examen en convocatoria ordinaria se le adjudicó una nota de 0. Se partió de una matriz de datos constituida por las observaciones de las notas obtenidas en las asignaturas cursadas (recogidas en columnas) por los distintos alumnos del mismo curso (recogidos en filas).

A partir de técnicas estadísticas de análisis multivariante se ha podido constatar la existencia de una estructura subyacente dentro del conjunto de todas las asignaturas cursadas por los alumnos de 1º de E4. Esta estructura queda plasmada a través de cuatro factores básicos e intrínsecos a la carrera seguida (Factor Cuantitativo, Factor Lingüístico Humanístico, Factor Empresarial y Factor 2º Idioma). Asimismo, a partir de estos factores y mediante un análisis de conglomerados se han encontrado ocho tipologías o patrones de comportamiento de los alumnos en relación con su rendimiento académico, a saber: buen rendimiento académico, situaciones atípicas, rendimientos académicos pésimos, rendimientos académicos malos, rendimientos académicos que destacan en alguna faceta y rendimiento académico normal.

La mayoría de los alemanes del Programa Alemán-Alemán se encuentran encuadrados dentro del conglomerado 1, esto es, el que mejores resultados presenta. El 50% de los alumnos del Programa Francés-Francés se encuentran dentro del conglomerado 8, es decir, el conglomerado donde los alumnos van aprobando aunque sin grandes alardes. El resto de los alumnos de este programa se encuentran dentro de otros conglomerados, en los que han obtenido resultados bastante pésimos, lo que provoca que en conjunto dicho programa sea el que salga peor parado, y proporcione los peores resultados. El programa Hispano-Americano se encuentra repartido entre los conglomerados 1, 5, 6 y 7 y la mayor parte de los alumnos del Programa Hispano-Alemán están en el conglomerado, seguido del conglomerado 1.

3.2.3 Estudio de validez predictiva de la PSU y comparación con el sistema PAA

Este trabajo fue realizado en el marco de la tesis de magíster de Sebastián Prado, de la Universidad de Chile el año 2008. Analizó la validez predictiva del Sistema PSU, en el ámbito de las carreras de ingeniería civil de dos universidades: la U. de Chile y la Pontificia U. Católica de Chile, mediante la estimación del rendimiento del primer año en la universidad. Además, a

partir de este estudio se estableció una comparación con el sistema PAA. Los datos utilizados correspondieron a los alumnos de primer año de las promociones desde el 2001 al 2006 ingresados a la carrera de ingeniería civil en las universidades señaladas.

Uno de los principales resultados obtenidos es que la validez predictiva de la PSU, para ingeniería civil en la PUC, es menor a la reportada en un estudio previo encargado por el Consejo de Rectores. Por otro lado, se encuentra que el porcentaje de varianza explicado es sensible a la nota de reprobación de los alumnos. Además, para esta casa de estudios, se observa que el número de alumnos que ingresaban a través de la PSU y reprobaban todos sus ramos, era mayor en un 2% a los alumnos ingresados vía PAA y que reprobaban todas las asignaturas de primer año. Los resultados presentados fueron robustos a controles por diferencias en la dificultad de los cursos que enfrentan los alumnos, variabilidad en la cantidad de alumnos admitidos año a año, diferencias en las escalas de transformación de puntajes entre ambos sistemas de selección. La implantación de los test de robustez mencionados y la utilización de información de los alumnos adicional a las pruebas (demográfica, socioeconómica y de los colegios de procedencia) en la estimación del rendimiento, constituyen las principales diferencias con el estudio del Consejo de Rectores citado.

4 Desarrollo del Proyecto

En este apartado se describirá lo realizado como desarrollo del proyecto. Todo el trabajo realizado para la consecución de los objetivos propuesto es detallado en las siguientes páginas. Para comenzar se describen las actividades realizadas en el marco del análisis inteligente de datos; continuando con el almacén de datos, el análisis ROLAP y finalizando con la minería de datos.

4.1 Integración de Datos

En este apartado se describen los procesos de ETL y de creación y llenado del almacén de datos realizados en el proyecto. Este proceso, como se describe en los capítulos anteriores, es iterativo e incremental, por lo tanto es siempre perfectible. Consiste en un proceso cíclico que no finaliza hasta concluir el proyecto. La herramienta utilizada de apoyo al proceso de integración de datos es Pentaho Data Integration (PDI).

4.1.1 Proceso de Extracción

Los datos que se han extraído para la investigación se describen en este apartado. Se han dividido en categorías o dimensiones, dependiendo si son de la procedencia del estudiante, de la batería de selección o del propio historial universitario. Además, debe usarse para cada estudiante un identificador único, que debe ser mantenido por igual por todas las dimensiones a las que se hacen referencia, o tablas que puedan generarse al extraer los datos. Para cada dimensión se describen los nombres de las variables y su significado. Cabe destacar que estos datos fueron los que finalmente se pudieron obtener desde la universidad, algunos datos no pudieron ser obtenidos tales como el ranking de egreso, el numero de preferencia....

4.1.1.1 Dimensión Procedencia

Estos datos son los de la procedencia del estudiante, y describen las condiciones que tenía el estudiante antes de ingresar a la Escuela. Los datos necesarios en esta dimensión son:

Quintil Representa el nivel socioeconómico del estudiante, clasificándolos según el ingreso de su grupo familiar.

TipoColegio Indica el establecimiento de egreso del estudiante, sea éste Particular, Subvencionado o Municipal.

Región Este dato representa la región donde se encuentra el establecimiento de origen del estudiante.

AnoEgreso Indica en qué año el estudiante egresó de Educación Media.

RamaEducacional Representa la modalidad de estudios del establecimiento de egreso del estudiante en Educación Media. Puede verse en detalle en la tabla 4.1.

AnoIngreso Para cada estudiante, se refiere al año al que ingresó a alguna de las carreras dictadas por la Escuela.

Tabla 4.1: Clasificación de la modalidad de estudios de Educación Media

Clasificación	Descripción
H1	Humanista Científico Diurno
H2	Humanista Científico Nocturno
H3	Humanista Científico – Validación de estudios
H4	Humanista Científico – Reconocimiento de estudios
T1	Técnico Profesional Comercial
T2	Técnico Profesional Industrial
T3	Técnico Profesional Servicios y Técnica
T4	Técnico Profesional Agrícola
T5	Técnico Profesional Marítima

NotaEMedia Este dato denota el promedio de las calificaciones del estudiante en la Educación Media. También se le conoce como NEM.

Educ_Padre Indica la escolaridad del padre del estudiante.

Educ_Madre Denota la escolaridad de la madre del alumno.

Sexo Representa el género del estudiante.

4.1.1.2 Dimensión Batería de Selección

Estos datos son propios del proceso de selección que es adoptado por la Escuela al pertenecer al Consejo de Rectores.

Promedio_PSU Puntaje promediado entre la PSU Lenguaje y la de Matemática.

PSU_Matemática, PSU_Lenguaje, PSU_HisyCsSociales, PSU_Ciencias Estas cuatro variables describen los puntajes obtenidos por el alumno en cada una de las pruebas específicas de la PSU.

4.1.1.3 Dimensión Universitaria

Los datos de esta dimensión son propios del paso por la Escuela del individuo, pretenden ser la base para los indicadores de desempeño que se pretenden confeccionar. Para cada una de las asignaturas cursadas por cada estudiante se genera un hecho que se guarda en la base de datos, con las siguientes variables. Para que quede más claro, una asignatura corresponde a la unidad perteneciente al currículo de cada carrera, en cambio un curso es una asignatura dictada en un semestre específico.

AnoCurso El año en que el estudiante cursa la asignatura actual

PeriodoCurso El semestre en que cursa la asignatura. Sea éste el primero o el segundo

CodAsignatura El código identificador único de cada asignatura.

NombreAsignatura Descripción de la asignatura.

CreditosAsignatura La cantidad de créditos especificados para cada asignatura.

NotaCurso La calificación obtenida por el estudiante al finalizar cada curso.

AprobacionCurso Denota si el alumno obtuvo, o no, la aprobación del curso.

SiglaAsignatura Describe, la mayoría de las veces, la Unidad Académica encargada de dictar la asignatura.

NumeroAsignatura Codificación interna que permite identificar distintas asignaturas de la misma Unidad Académica.

Dado que no es posible tener acceso directo a los datos de la universidad el proceso de extracción realizado en este proyecto es el desde el origen entregado por la misma universidad. Este origen consta de una hoja de cálculo incluyendo los distintos datos ya extraídos de las bases de datos de la institución académica. Aunque en esta investigación no se realiza un proceso de extracción de diferentes fuentes de información no significa que este proceso no esté presente. La extracción entonces queda enmarcada en obtener los datos desde la hoja de cálculo.

Para lograr la obtención desde el archivo se debió realizar una puesta en escena. Esta actividad se realizó sobre una base de datos MySQL llamada «test». Desde esta base de datos también se realizó el proceso de transformación. El diagrama de entidad-relación mejorado (EER) de esta base de datos se puede ver en la figura 4.2. La elección de este método se basa en que es de mayor facilidad transformar los datos en una base de datos que en la misma u otra hoja de cálculo. Nótese que esta actividad es un proceso de ETL en sí, la extracción desde el archivo, la transformación filtrando filas y la carga en las tablas de datos de transición.

La primera fase de la extracción se realiza separando la data de los estudiantes dependiendo de la carrera a la que están adscritos. Esta fase puede verse descrita en la figura 4.1. En la misma puede verse que la herramienta permite seleccionar los datos que se desean cargar. La segunda etapa de extracción se hace desde la misma base de datos para el proceso de transformación.

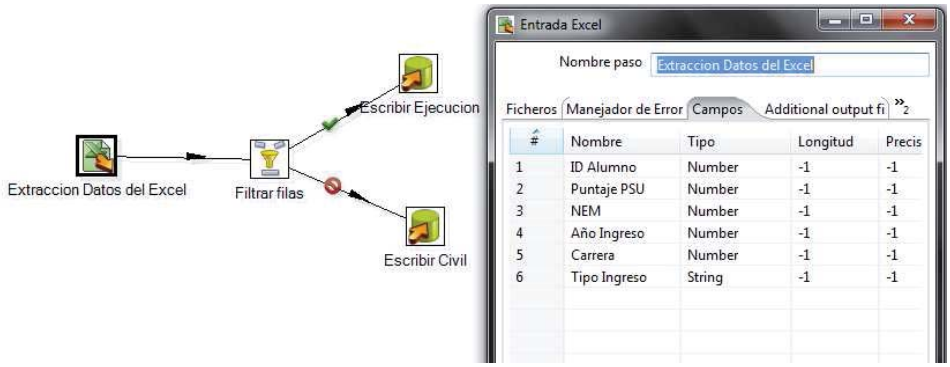

Figura 4.1: Proceso de Extracción

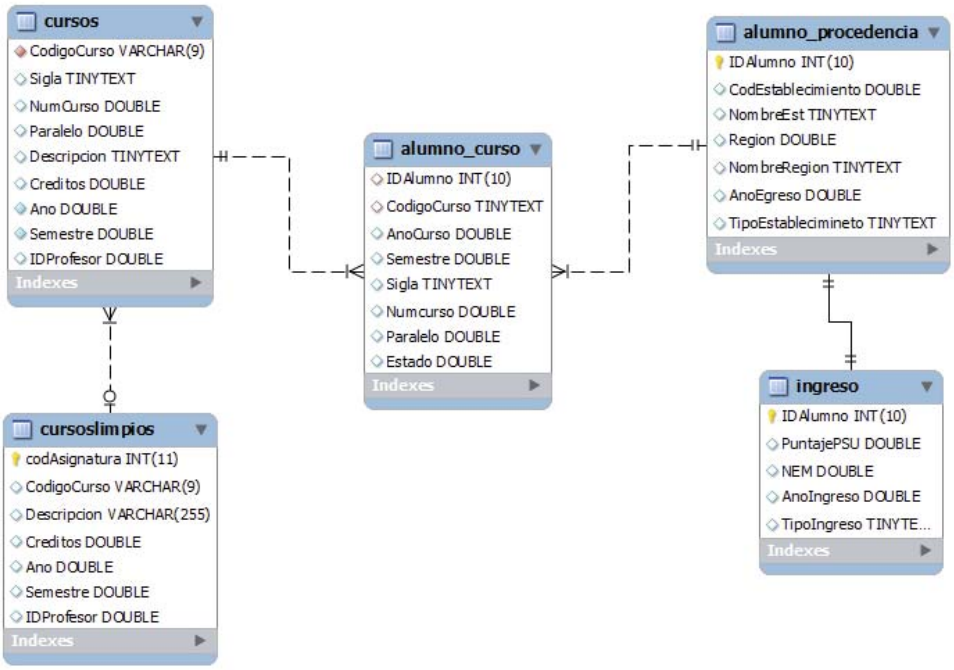

Figura 4.2: Diagrama EER de la etapa de puesta en escena

4.1.2 Proceso de Transformación

Este proceso es un conjunto de etapas, que se describen en el marco conceptual. Para la transformación de los datos este proceso se apoya en las actividades de validación, limpieza, decodificación e identificación. A continuación se describen en detalle cada una de estas actividades realizadas.

- Validación: Al realizar un análisis visual de los datos se encontraron distintos elementos que no eran válidos. Se hizo un filtro dejando de lado por el momento las asignaturas que se están cursando al momento de la extracción y que por lo tanto no tienen el dato de terminación (aprobación o reprobación), además de separar también a los estudiantes ingresados el año 2010, ya que poseen solamente un año de registros universitarios. Algunas columnas que no aportaban información relevante fueron dejadas de lado del almacén de datos, como por ejemplo el nombre del colegio, o el tipo de prueba que dio para ingresar (ya que la totalidad ingresó vía PSU), el ingreso bruto familiar, entre otras.

- Limpieza: Se realizó un análisis para verificar si existían elementos que debían ser corregidos. Se encontraron registros que no tenían finalización ya que se encontraban en curso al momento de la extracción.

- Decodificación: La columna de región del establecimiento de origen fue decodificada, y pasó de ser numérica arábiga a romana para una mejor comprensión del usuario final.

- Identificación: Para cada asignatura cursada por cada estudiante se genera un código identificador autoincremental para la tabla de hechos del almacén de datos. También para cada asignatura dictada en momentos distintos se realiza lo mismo.

También es parte del proceso de transformación la unión en dimensiones de los diferentes elementos. El almacén de datos, que es explicado en el apartado siguiente, consta de dos dimensiones y dos tablas de hechos. La dimensión del alumno agrupa todos los datos del ingreso y el origen del estudiante, mientras la dimensión curso las asignaturas y sus componentes. La transformación de estos datos se hizo con ayuda tanto de PDI como de consultas SQL para casos específicos.

4.1.3 Proceso de Carga

Este proceso se realizó en la manera de *job*, ya que semestralmente debe cargarse nuevamente el almacén de datos con los nuevos elementos que se generan en la Escuela. Como puede verse en la figura 4.3, se llena el almacén de datos con las dimensiones establecidas y luego las tablas de hechos. Para el llenado de la ultima tabla de hechos se hace una ETL desde la tabla de hechos de rendimiento, ya que la data del primer año es un subconjunto de la data total.

4.2 Almacén de Datos

El almacén de datos consiste en una base de datos hecha con el propósito de realizar con ella un análisis OLAP. Se diseña antes de aplicarse el proceso de carga. Para el diseño del almacén se utilizó la aplicación MySQL Workbench, cuyo desarrollador es la reconocida empresa MySQL. El diseño debe tener en cuenta el tipo de análisis a considerar (ROLAP, MOLAP o híbrido) y el esquema de base de datos que se utilizará (estrella, copo de nieve, etc.).

El modelado del almacén de datos, como puede verse en su diagrama EER de la figura 4.4, constó de dos dimensiones y dos tablas de hechos. El análisis que se realizará es ROLAP ya que la base de datos es del tipo relacional (MySQL). El tipo de esquema del almacén de datos es estrella. El esquema se escogió para una mayor facilidad al momento de construir los cubos en la etapa de análisis.

En la dimensión alumno se consolidaron los datos del establecimiento del cual egresó, junto con los datos de la procedencia del estudiante. El identificador de esta dimensión es un ID que desde la universidad se asignó en vez del RUT por un tema de confidencialidad. Se nombraron los atributos de la dimensión con el objetivo de que sea lo mas descriptivo posible.

En la dimensión curso se consolidó toda la información de las asignaturas dictadas por la escuela en todos los semestres. Se estudió la posibilidad de realizar aquí otra tabla de hechos con las asignaturas dictadas y una dimensión tiempo, pero se optó por consolidar todas las asignaturas en la dimensión ya que no existen problemas de rendimiento. La columna identificadora se construyó en la etapa de identificación del proceso de transformación

Figura 4.3: Job de carga del almacén de datos

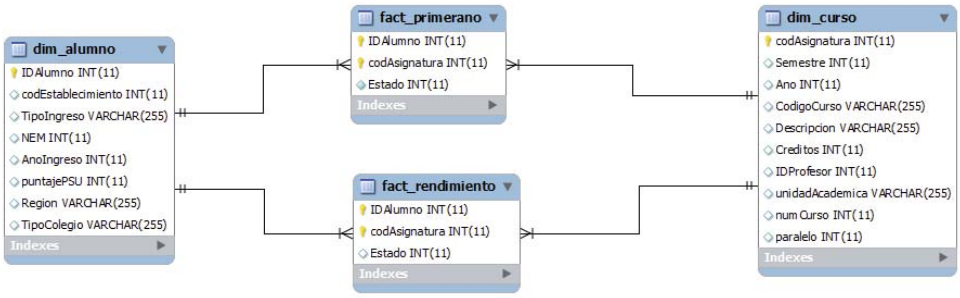

Figura 4.4: Diagrama EER del almacén de datos

En la tabla de hechos de rendimiento se registran todas las asignaturas cursadas por los estudiantes. La medida estado, del tipo binaria, muestra si el estudiante aprobó el curso o no. Lo mismo para la tabla de hechos del primer año, esta tabla es un subconjunto de la tabla de hechos de rendimiento. El llenado se realizó en la etapa de transformación comparando si la asignatura cursada era del mismo año del ingreso.

El diseño del almacén de datos se realiza con el objetivo de que los procesos ETL siguientes no modifiquen sustancialmente el almacén. Es por esto que las dimensiones deben ser creadas con el propósito de que se modifiquen o se agreguen fácilmente registros en el tiempo. Con el almacén ya listo puede realizarse el análisis, en este caso ROLAP. En el apartado siguiente se describirá este análisis y los pasos que se realizaron para lograrlo.

4.3 Análisis de los Datos

4.3.1 Análisis Descriptivo de los Datos

El análisis descriptivo se realiza a través de la definición de métricas estadísticas que permiten controlar la presencia de posibles errores o datos atípicos que hayan resultado del proceso de ETL anterior. Además, permite hacerse una idea de la forma de los datos, su distribución y sus parámetros de centralidad (por ejemplo la media); como también sus parámetros de dispersión, tales como la desviación estándar.

Para esta fase se utilizó como software de apoyo la herramienta IBM SPSS Statistics, cuyos módulos permiten realizar un análisis rápido y sin errores. Se analizó, en un comienzo, los datos de ingreso de los estudiantes para después realizar una examinación de los datos universitarios de los alumnos. Para comenzar se realizó un análisis de frecuencias, se calcularon las variables de centralidad y dispersión para las variables cuantitativas y se generó una tabla de frecuencias para las cualitativas.

4.3.1.1 Análisis de los Datos de Ingreso

Se realizó un análisis descriptivo de los datos de ingreso para las cohortes comprendidas entre el año 2006 al 2009. La tabla 4.2 muestra el resultado del cálculo de las parámetros de centralidad y dispersión para las variables cuantitativas.

Por una parte, las variables cuantitativas se comportan de una manera similar. Las NEM tienen una media de 5.99 aproximadamente con una desviación estándar cercana a 0.34, su dato

central está cercano a la media y es igual a su moda; su distribución es cercana a la simetría de la normal y es un poco más achatada que ésta; no presenta datos perdidos ni atípicos (su mínimo y máximo son coherentes). Las pruebas en específico presentan distribuciones similares, aunque las pruebas optativas (Historia y Ciencias) tienen una gran cantidad de datos perdidos (Historia con más del 50 % del total), ademas de presentar asimetría hacia la derecha. En la PSU de ciencias se presenta un dato atípico: un puntaje mínimo de 270, aunque esto puede haber sucedido en la realidad, dada la baja ponderación de esta prueba al ingresar a la carrera. Las pruebas presentan una distribución menos normal que las NEM pero aún así se comportan de esta manera, que por cierto es esperada.

Por otra parte, las variables cualitativas se presentan en tablas de frecuencias para cada una de ellas. Por cuestión de síntesis se describirán las más importantes; éstas son el ano de ingreso, el tipo de colegio, la región, el sexo y el quintil. Las tablas de frecuencias se detallan en las tablas 4.44.34.54.64.7 y 4.8. De estas tablas se puede ver que la cantidad de estudiantes que ingresan a la carrera ha ido en aumento desde el 2006 al 2009, también que la mayoría de los alumnos provienen de colegios particulares-subvencionados (con un 57,6 % del total), de la quinta región (un 50,8 % del total) y son hombres (88 % del universo).

4.3.1.2 Análisis de los Datos Académicos

Además del análisis descriptivo para los datos de ingreso, se realizó un análisis de las variables académicas, éstas son las notas de los cursos, las asignaturas dictadas y el estado final de la asignatura. La examinación se realizó de la misma manera que la de los datos de ingreso, separando las variables cualitativas de las cuantitativas. Para el caso de la nota del curso, se optó por estandarizar los datos por cada curso dictado, llevando a las notas de la asignatura a tener una distribución normal; esto es, con media cero y desviación estándar uno. Esto último permite tener una nota relativa a la media y dispersión por cada curso, ya que existen factores específicos que afectan cada vez que una asignatura es tomada por el estudiante (profesor, aprendizaje relativo, sucesos imprevistos, etc). La forma de llevar estos datos a la distribución Z (normal) es la clásica mostrada en la ecuacion 4.3.1.

$$Z(x_i) = \frac{x_i - \bar{x}}{\sigma} \tag{4.3.1}$$

Como puede verse en la tabla 4.9 la nota de los cursos presenta una media cercana al 3.41, un dato central de 3.5, una desviación estándar de aproximadamente 1.24, una asimetria hacia la derecha de 0.148 y una forma mas puntiaguda que la distribución normal; no presenta datos erróneos ya que el mínimo y el máximo se encuentran dentro de los valores posibles. Por otra parte, al transformar la distribución a Z presenta una mayor asimetría pero una menor curtosis; se disminuye tambien la varianza y la desviación estándar, tal como se ve en la figura 4.6.

Cabe destacar que la distribución original presenta una moda con una frecuencia totalmente fuera de lo común en la nota límite de aprobación (ver figura 4.5), esto supone que existen mecanismos aparte de las pruebas comunes, que permiten aprobar los cursos con la nota mínima, lo cual distorsionará el modelo predictivo.

Para las variables cualitativas se confeccionaron tablas de frecuencias. La tabla 4.10 mues-

Tabla 4.2: Variables de centralidad y dispersión de los datos de ingreso, cohortes 2006-2009

		NotaEMedia	PSU_Lenguaje	PSU_Matematica
N	Válidos	309	309	309
	Perdidos	0	0	0
Media		5,9961	605,51	641,11
Error típ. de la media		,01947	3,067	2,114
Mediana		6,0000	602,00	640,00
Moda		6,00	581	652
Desv. típ.		,34224	53,911	37,167
Varianza		,117	2906,387	1381,367
Asimetría		-,108	,522	,404
Error típ. de asimetría		,139	,139	,139
Curtosis		-,263	,509	,458
Error típ. de curtosis		,276	,276	,276
Rango		1,70	329	213
Mínimo		5,10	458	550
Máximo		6,80	787	763
Percentiles	25	5,8000	569,00	616,00
	50	6,0000	602,00	640,00
	75	6,2000	638,00	661,00

Tabla 4.3: Variables de centralidad y dispersión de los datos de ingreso, cohortes 2006-2009 (cont.)

		Promedio_PSU	PSU_HisyCsSociales	PSU_Ciencias
N	Válidos	309	153	272
	Perdidos	0	156	37
Media		623,3123	589,91	583,57
Error típ. de la media		1,84591	5,266	3,073
Mediana		619,0000	589,00	587,00
Moda		616,00	613	609
Desv. típ.		32,44807	65,133	50,688
Varianza		1052,877	4242,294	2569,309
Asimetría		,706	-,009	-,788
Error típ. de asimetría		,139	,196	,148
Curtosis		,735	-,210	4,956
Error típ. de curtosis		,276	,390	,294
Rango		187,00	329	496
Mínimo		557,50	428	271
Máximo		744,50	757	767
Percentiles	25	600,0000	542,00	552,00
	50	619,0000	589,00	587,00
	75	643,0000	633,00	613,00

Tabla 4.4: Tabla de frecuencia del año de ingreso de los alumnos

Año	Frecuencia	Porcentaje
2006	74	23,9
2007	73	23,6
2008	80	25,9
2009	82	26,5
Total	309	100,0

Tabla 4.5: Tabla de frecuencia del tipo de colegio de los estudiantes

Tipo de Colegio	Frecuencia	Porcentaje
MUNICIPAL	67	21,7
PARTICULAR	64	20,7
SUBVENCIONADO	178	57,6
Total	309	100,0

Tabla 4.6: Tabla de frecuencia de la región de origen de los estudiantes

Región	Frecuencia	Porcentaje
1	11	3,6
10	7	2,3
11	1	,3
12	3	1,0
13	62	20,1
15	2	,6
2	2	,6
3	4	1,3
4	10	3,2
5	157	50,8
6	40	12,9
7	6	1,9
8	3	1,0
9	1	,3
Total	309	100,0

Tabla 4.7: Tabla de frecuencia del sexo de los estudiantes

Sexo	Frecuencia	Porcentaje
HOMBRE	272	88,0
MUJER	37	12,0
Total	309	100,0

Tabla 4.8: Tabla de frecuencia del quintil socioeconómico de los estudiantes

Quintil	Frecuencia	Porcentaje
1	53	17,2
2	61	19,7
3	45	14,6
4	47	15,2
5	103	33,3

Tabla 4.9: Variables de centralidad y dispersión de los datos académicos, cohortes 2006-2009

		NotaCurso	NotaCurso_Stand
N	Válidos	2695	2695
	Perdidos	0	0
Media		3,4117	,0000
Error típ. de la media		,02389	,01907
Mediana		3,5000	,0579
Moda		4,00	,48
Desv. típ.		1,24023	,98993
Varianza		1,538	,980
Asimetría		-,148	-,222
Error típ. de asimetría		,047	,047
Curtosis		-,813	-,443
Error típ. de curtosis		,094	,094
Rango		5,80	6,11
Mínimo		1,00	-3,16
Máximo		6,80	2,95
Percentiles	25	2,4000	-,7338
	50	3,5000	,0579
	75	4,3000	,7365

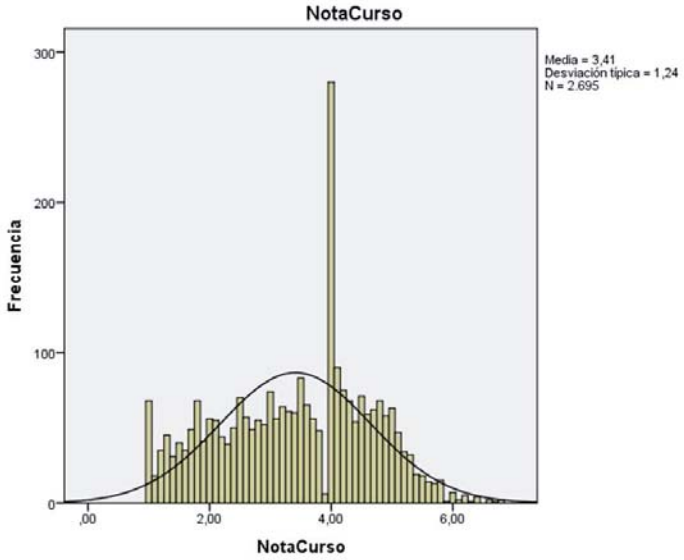

Figura 4.5: Histograma de las notas de los cursos dictados entre 2006 y 2009

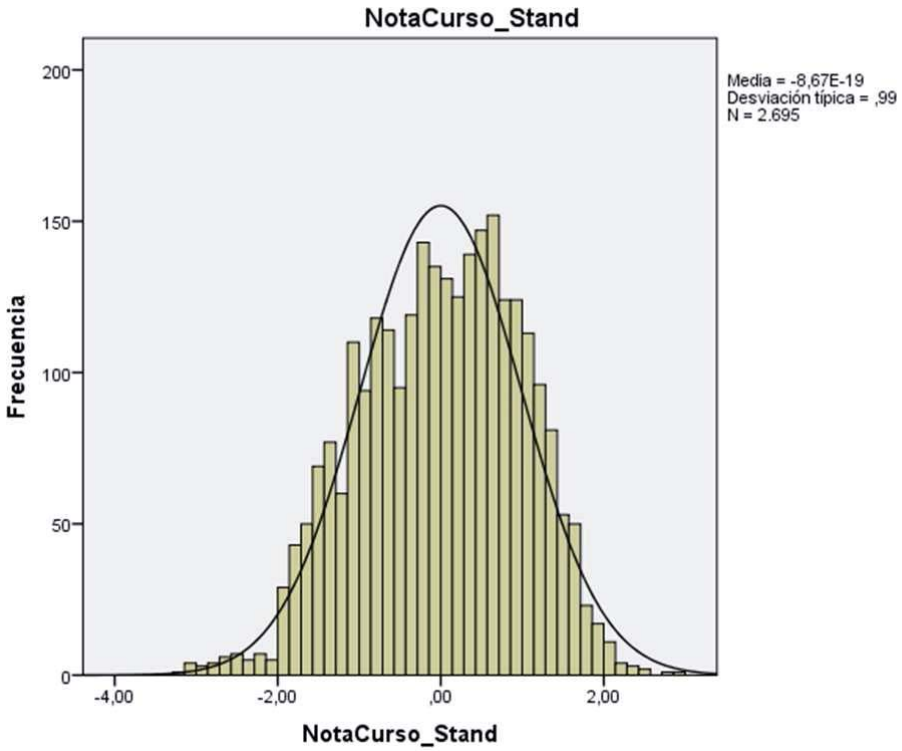

Figura 4.6: Histograma de las notas estandarizadas de los cursos dictados entre 2006 y 2009

tra las asignaturas que se dictan el primer año de la carrera y la cantidad de inscripciones que han realizado para cada asignatura. La tabla 4.11 muestra el estado final de los cursos inscritos por los estudiantes, tanto si aprobaron (A) o reprobaron (R). Finalmente, la tabla 4.12 muestra la cantidad de inscripciones totales por área de estudio; esto es, las asignaturas de Ciencias, Matemática o propias de la Informática.

Las tablas de frecuencia muestran que la mitad de las inscripciones son en cursos del área matemática, influenciado porque existen más asignaturas de este tipo en el currículo. Si se estudia por asignatura, la que tiene mayor inscripcion es la de Cálculo Diferencial; tomando como referencia que todos los que ingresan cursan Introducción a la Ingeniería Informática es posible ver que las asignaturas que más reprueban son las de matemática, e incluso la deserción es tanta que un porcentaje menor de estudiantes inscriben las asignaturas del segundo semestre, tales como Cálculo Integral y Series y Física General Mecánica. Ésto último puede verse tambien reflejado en la tasa de reprobación total que llega al 56,8%.

4.3.2 Cubos Confeccionados

El proceso de creación de los cubos ROLAP se realizó utilizando la herramienta de Pentaho llamada Schema Workbench (PSW). Esta herramienta permite crear las expresiones MultiDimensional eXpressions (MDX) en modo gráfico. MDX es un lenguaje de consultas OLAP. Las expresiones MDX interactúan con el motor de base de datos y a través de metadatos realiza consultas SQL al almacén de datos.

Debido a que en un principio el almacén de datos fue diseñado en un esquema estrella, los cubos ROLAP confeccionados tienen una relación con las dimensiones del almacén planteadas. Para el caso específico de este proyecto se confeccionó una totalidad de cuatro cubos. Dos cubos corresponden a las dos dimensiones y los otros dos cubos a las tablas de hechos. Los cubos de las dimensiones fueron creados con la finalidad de realizar cubos virtuales más adelante en el proyecto.

Para facilitar el desarrollo de los cubos se optó por la modalidad de crear dimensiones compartidas. Este tipo de dimensiones permite que puedan ser usadas en más de un cubo a la vez. Si no se utilizara esta modalidad, se tendrían que crear dimensiones para cada cubo, pero al tener dimensiones comunes se debe evitar el re-trabajo. Cuando se usa una dimensión compartida en

Tabla 4.10: Tabla de frecuencia de los cursos inscritos por los estudiantes

Asignatura	Frecuencia	Porcentaje
ALGEBRA	530	19,7
CALCULO DIFERENCIAL	532	19,7
CALCULO INTEGRAL Y SERIES	286	10,6
FISICA GENERAL MECANICA	243	9,0
FUNDAMENTOS DE PROGRAMACION	366	13,6
FUNDAMENTOS DE QUIMICA	322	11,9
INTRODUCCION A LA INGENIERIA INFORMATICA	416	15,4
Total	2695	100,0

Tabla 4.11: Tabla de frecuencia para el estado de aprobación de cada curso

Estado	Frecuencia	Porcentaje
A	1165	43,2
R	1530	56,8
Total	2695	100,0

Tabla 4.12: Tabla de frecuencia por área de estudio

	Frecuencia	Porcentaje
CS	565	21,0
INF	782	29,0
MAT	1348	50,0
Total	2695	100,0

un cubo se denomina *dimension usage*. Las dimensiones compartidas son: alumno, colegio, cohorte, región, curso y período.

Se creó un cubo para cada dimensión: «ingreso» y «cursos», y un cubo para cada tabla de hechos: «rendimiento» y «primerano». Para el caso del cubo «ingreso» las dimensiones son: alumno, colegio, región y cohorte. Mientras que en el cubo «cursos» son dos: curso y período. El cubo «rendimiento» y «primerano» son similares, ya que uno es un subconjunto del otro, solo difiriendo en que en el cubo «rendimiento» se utiliza la dimensión período, que es inútil en el cubo del primer año porque las asignaturas son siempre del mismo período de la cohorte. Estos últimos dos cubos tienen las siguientes dimensiones: colegio, cohorte, región y asignatura.

Las métricas o medidas utilizadas en los cubos muestran los datos cuantitativos que se agregan para obtener información. Para todos los cubos se utilizó la medida de aprobación porcentual, denominada por el autor como Tasa de Rendimiento Porcentual (TRP), definida en la ecuación 4.3.2. El en caso de los cubos «rendimiento», «primerano» e «ingreso» se utiliza la cantidad de alumnos que contiene la agregación, con el propósito de verificar que las mismas se estén realizando de manera correcta. También se utilizó la desviación estándar tanto absoluta (ecuación 4.3.3) como relativa (RSD) (ecuación 4.3.4)en algunos cubos como métrica calculada para obtener la precisión de los datos ponderados. Todo el detalle de los cubos puede verse en la figura 4.7.

$$TRP = \frac{AsignaturasAprobadas}{AsignaturasInscritas} * 100 \tag{4.3.2}$$

$$s = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}} \tag{4.3.3}$$

$$RSD = \frac{s}{\overline{x}} = \frac{\sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}}}{\overline{x}} \tag{4.3.4}$$

4.3.3 Análisis ROLAP

Se realizó un análisis ROLAP para tres de los cuatro cubos: «ingreso», «primerano» y «rendimiento». El cuarto cubo «cursos» no se analizó mayormente ya que la única métrica posible fue los créditos que posee cada asignatura. A continuación se desagregará el análisis para cada uno de los tres cubos estudiados.

Los gráficos se realizaron desde los datos normalizados. La forma de normalizar los datos fue la de la ecuación 4.3.5, denominada normalización min-max. Con esta fórmula los datos quedan distribuidos en el intervalo [0:1], permitiendo la comparación entre dos grupos de datos con diferente escalas, como es el caso del puntaje PSU con respecto a las NEM.

$$norm(x_i) = \frac{x_i - min(x_0 : x_{n-1})}{max(x_0 : x_{n-1}) - min(x_0 : x_{n-1})} \tag{4.3.5}$$

4.3.3.1 Exploración del Cubo Ingreso

Lo primero que se desea analizar al ver el cubo de ingreso es la relación existente entre el puntaje PSU y las NEM. Para la mayoría de la gente resultaría normal encontrar una relación positiva entre estas dos variables. El sentido común lleva a pensar que si un estudiante tiene notas de enseñanza media mas altas que el promedio también tendría un puntaje PSU por sobre el promedio. El análisis, realizado a los más de 500 registros de estudiantes que han pasado por la carrera de Ingeniería Civil Informática demuestra lo contrario. En promedio, mientras mas alto es el puntaje PSU mas bajas son las NEM.

Si se explora mediante las diferentes dimensiones del cubo también se ve la tendencia inversa que presentan las dos variables. La tabla 4.13 muestra la exploración del cubo mediante la dimensión cohorte. Se puede notar que la cohorte de 2004 fue la que obtuvo un puntaje promedio PSU más alto y también unas NEM más altas, lo cual representa un dato atípico ya que desde el 2005 hacia adelante el puntaje PSU ha mostrado una tendencia positiva mientras las NEM la contraria.

La exploración ahora mediante la dimensión región puede ser de ayuda para el modelo predictivo posterior. La tabla 4.14 muestra que por ejemplo la IX región presenta los puntajes PSU mas altos y unas NEM promedio. Las regiones mas extremas de Chile tienen los puntajes

Tabla 4.13: Exploración Cubo Ingreso por cohorte

Cohorte	Puntaje PSU	NEM	Alumnos	RSD. PSU	RSD.NEM
All cohortes	623	60	530	5,69%	5,64%
2004	639	61	67	6,61%	4,96%
2005	608	61	77	6,19%	5,20%
2006	621	60	73	4,70%	5,41%
2007	622	60	73	5,14%	6,28%
2008	619	60	80	5,24%	4,98%
2009	633	59	82	5,46%	6,02%
2010	621	60	78	5,20%	6,08%

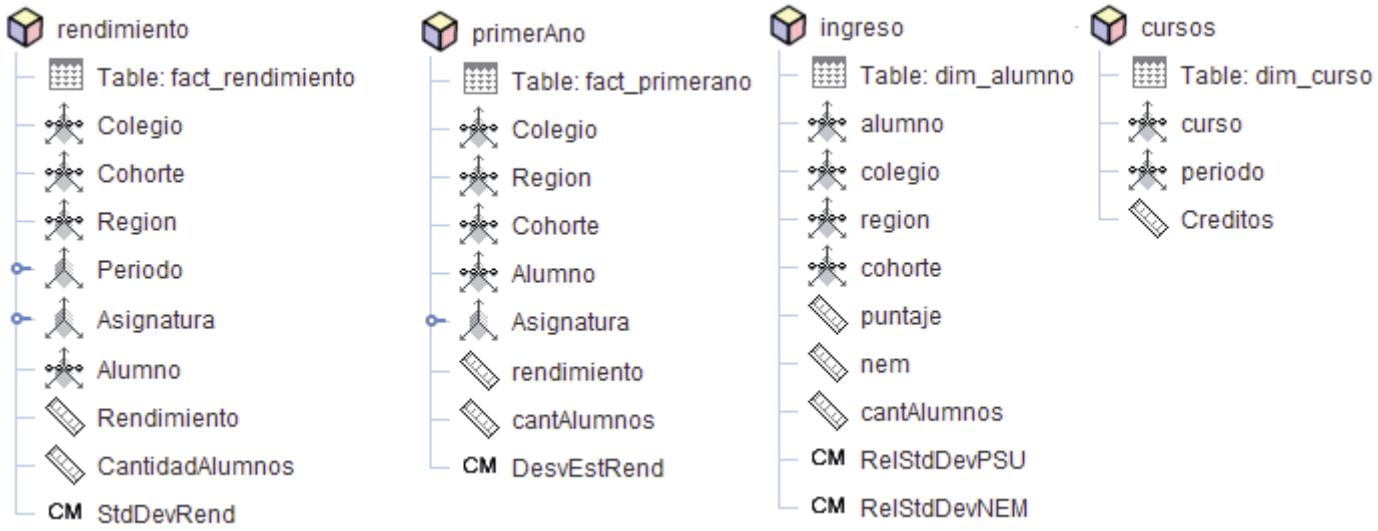

Figura 4.7: Detalle de los cubos confeccionados

PSU mas bajos, pero las NEM están por sobre el promedio de las regiones, aunque la poca cantidad de estudiantes de esas regiones no permitan generalizar con propiedad.

Ahora bien, si la exploración se hace mediante el tipo de colegio, esto es particular, subvencionado o municipal, como se ve en la figura 4.15 se puede notar que los que egresaron de un colegio privado tienen mayor puntaje PSU en promedio pero menores NEM. En el otro extremos están los que llegan desde un colegio municipal con un puntaje PSU por debajo de la media y unas NEM por sobre ésta. Pero hay que hacer hincapié en que la variabilidad que presentan los egresados de colegios particulares es más alta que en los otros tipos de colegios tanto en PSU como en NEM. La cantidad de estudiantes de los subvencionados es casi la misma que los de municipal y particular juntos y presenta la más alta estabilidad en la variabilidad de los puntajes, con estos dos bordeando el promedio.

4.3.3.2 Exploración del Cubo de Primer Año

Se hará una exploración en cuatro dimensiones. Las dimensiones son el tipo de colegio, la región de origen, la cohorte del estudiante y el tipo de asignatura cursada en el primer año. La métrica de este cubo es el rendimiento promedio porcentual, definido como la razón entre las asignaturas aprobadas con respecto a las inscritas. Se utilizó como medida de precisión la desviación estándar absoluta *s* (SD).

El tipo de asignatura está definida como la sigla que comienza su codificación. Esta sigla es referente a la unidad académica a la cual la asignatura pertenece. Si es diferente a ICI entonces la asignatura es de servicio, es decir, que otra unidad académica le provee la asignatura a los estudiantes de la Escuela.

Para la exploración del tipo de colegio, que puede verse en la tabla 4.16, se ve que los estudiantes que provienen de colegios particulares poseen un mayor rendimiento promedio que los demás, con una variabilidad mas baja también. Los de colegios municipales tienen un rendi-

Tabla 4.14: Exploración Cubo Ingreso por región

Region	Puntaje PSU	NEM	Alumnos	RSD.PSU	RSD.NEM
All regions	623	60	530	5,69%	5,64%
IX	645	60	3	5,43%	0,96%
X	638	61	18	6,17%	5,65%
VIII	637	60	5	4,83%	4,83%
II	630	60	7	9,21%	5,04%
IV	626	60	22	4,67%	5,30%
V	624	61	284	5,89%	5,34%
RM	623	58	89	5,17%	6,25%
XI	623	58	2	1,70%	3,69%
I	619	61	14	7,08%	4,12%
VI	616	60	60	5,02%	5,28%
III	613	62	7	6,60%	4,61%
VII	613	60	12	5,28%	5,25%
XII	613	62	5	4,07%	6,55%
XV	596	61	2	2,49%	3,51%

Tabla 4.15: Exploración Cubo Ingreso por tipo de colegio

Colegio	Puntaje PSU	NEM	Alumnos	RSD.PSU	RSD.NEM
All colegios	623	60	530	5,69%	5,64%
Municipal	613	61	118	5,40%	5,67%
Subvencionado	621	60	284	5,22%	5,30%
Pagado	638	59	128	6,21%	6,07%

miento mas bajo que el promedio y además el indicador posee una precisión mayor.

Al explorar mediante la dimensión región (tabla 4.17) se puede ver que las regiones X, II y IX son las que mejor TRP tienen, aunque poseen poca cantidad de estudiantes. Las regiones RM, XV, I y XI son las que tienen el rendimiento mas bajo. Da el caso que en los últimos lugares están las regiones mas extremas del país, pero la cantidad de estudiantes no es suficiente para generalizar. La RM es una de las regiones cercanas a la Universidad, por lo tanto con mas estudiantes, y posee un rendimiento visiblemente mas bajo que el promedio de las regiones.

Explorando por cohorte puede notarse que el rendimiento en el primer año ha ido en una tendencia a la baja a través de los años. Uno de los años mas bajos en rendimiento fue el 2009, seguido por el 2006. Además puede notarse que las vacantes han ido en aumento. El valor del r_{xy} entre el RPP y la cantidad de estudiantes es de -0.6607, lo cual sugiere una relación inversa un poco fuerte.

Mediante la exploración de la dimensión tipo de asignatura pueden verse en la figura los rendimientos dependiendo de la unidad académica que dicta la asignatura. Existen asignaturas en esta vista que no corresponden a la malla del primer año de la carrera. Lo último podría estar explicado por estudiantes que provienen de otras casas de estudios y realizan convalidaciones con lo que podrían inscribir asignaturas de tercer o cuatro semestre en adelante. Otra de las causas es la anticipación de la inscripción de algunas asignaturas con respecto a la malla curricular. Las asignaturas de este tipo pueden notarse por la poca cantidad de estudiantes que las inscriben, ademas de pertenecer a unidades académicas aparte de ICI, FIS, MAT, y QUI.

Se puede observar en los datos que los estudiantes que inscriben ramos por adelantado a la malla curricular logran mejores resultados que el promedio llegando a alcanzar un 85,71% en EIE. La única excepción es ICM que posee incluso una tasa menor que la de las asignaturas de la carrera propiamente tal, pero más alta que el promedio. Las áreas de estudio con mayor reprobación son las de matemática, pero su variabilidad es mas alta, al igual que la de FIS.

Otra cosa que puede verse en la tabla es que aproximadamente un tercio de los estudiantes que ingresan logran inscribir los ramos de FIS, esto es debido a que el prerequisito es una asignatura de MAT, la cual tiene la menor tasa de aprobación

4.3.3.3 Exploración del Cubo de Rendimiento

Este cubo contiene todos los hechos, es decir, todas las asignaturas inscritas por todos los estudiantes y su estado final, sea éste aprobación o reprobación. Se realizarán exploraciones de este cubo por cada dimensión, tal como fue realizado con los dos anteriores. Para cada dimensión

Tabla 4.16: Exploración Cubo Primer Año por tipo de colegio

Colegio	Rendimiento	Alumnos	SD.Rendimiento
All colegios	45,14%	452	29,35%
Pagado	50,06%	115	27,55%
Subvencionado	44,32%	243	30,09%
Municipal	41,19%	94	28,95%

Tabla 4.17: Exploración Cubo Primer Año por región

Region	Rendimiento	Alumnos	SD.Rendimiento
All regions	45,14%	452	29,35%
X	65,31%	14	23,03%
II	62,16%	5	17,78%
IX	61,11%	3	46,30%
XII	57,58%	5	33,80%
III	51,35%	6	27,45%
VIII	50,00%	5	22,61%
V	46,41%	242	28,77%
IV	46,23%	17	36,57%
VI	44,55%	51	30,20%
VII	44,05%	12	31,84%
RM	38,09%	75	27,64%
XV	35,71%	2	10,10%
I	28,09%	13	26,51%
XI	7,14%	2	10,10%

Tabla 4.18: Exploración Cubo Primer Año por cohorte

Cohorte	Rendimiento	Alumnos	SD.Rendimiento
All cohortes	45,14%	452	29,35%
2004	52,75%	67	30,25%
2005	49,03%	77	29,13%
2006	39,20%	73	29,96%
2007	49,90%	73	30,36%
2008	43,47%	80	29,58%
2009	37,97%	82	24,36%

Tabla 4.19: Exploración Cubo Primer Año por tipo de asignatura

Asignatura	Rendimiento	Alumnos	SD.Rendimiento
All Asignaturas	45,14%	452	29,35%
EIE	85,71%	4	25,00%
ICA	77,78%	9	44,10%
EII	63,64%	9	48,59%
QUI	54,99%	391	49,81%
ICI	53,74%	452	32,24%
ICM	52,94%	11	41,56%
FIS	45,45%	154	49,96%
MAT	37,29%	450	34,19%

se mostrará una tabla con la exploración correspondiente.

Si se explora mediante la dimensión tipo de colegio, puede verse que los estudiantes que provienen de un colegio municipal tienen un rendimiento promedio mas bajo que los de otros tipos de colegio. Los que provienen de particulares poseen un rendimiento bastante mayor que el promedio y su variabilidad es mas baja que la media.

Al explorar la dimensión cohorte es posible visualizar que existe una relación del rendimiento promedio con los años. Esto es debido a que en la instantánea de estos datos (año 2010) los estudiantes que ingresaron el 2004 tienen más asignaturas inscritas que los de 2009. La única cohorte que rompe esta tendencia es la del 2006, que presenta casi 10 puntos menos que la del 2007 habiendo tenido un año más para recuperarse. La variabilidad de todas las cohortes no se aleja de la media.

Al hacer el análisis desagregando por región se puede ver que los estudiantes que provienen de la novena región son los que tienen la tasa mas alta de aprobación, seguidos por los de la décima. Las regiones con más bajos rendimientos son la primera, decimoquinta y undécima. La tónica es que las regiones mantienen sus puestos en el *ranking* comparando con el de primer año, salvo contadas excepciones, la VI que sube y la II que estrepitosamente baja. Cabe destacar que la IX región tiene una alta variabilidad y pocos estudiantes, lo cual presume que un estudiante de esos tres es el que hace subir el promedio.

Para las siguientes dos exploraciones ha debido dejarse de lado el calculo de la desviación estándar ya que la cantidad de registros impide el cálculo con la herramienta utilizada (Mondrian). La herramienta no tiene la capacidad de manejar en memoria tantos cálculos y desborda el búfer con lo que no se obtienen los resultados. Sin embargo, al eliminar esa columna es posible explorar las dimensiones tanto de tipo de curso como de periodo de las asignaturas.

Al examinar la tabla 4.23 se puede apreciar que existen tipos de asignaturas que poseen un rendimiento promedio elevado. Tal es el caso de derecho y comercial (tanto COM como ICA). En un segundo grupo están las asignaturas que se imparten en el mismo edificio donde se ubica la escuela (EII e ICI) y el ultimo grupo que son lo de las ciencias básicas que tienen un rendimiento inferior a la media. Cabe destacar aquí que las asignaturas que se encuentran en el grupo de mayor rendimiento coinciden con que se dictan en los últimos años de la malla curricular, donde el rendimiento comienza a subir notoriamente.

La ultima dimensión se refiere al período donde de cursa cada asignatura. Existe una tendencia positiva, cada año el rendimiento promedio se eleva con respecto al anterior. Sin embargo,

Tabla 4.20: Exploración Cubo Rendimiento por tipo de colegio

Tipo de Colegio	Rendimiento	Alumnos	SD.Rendimiento
All colegios	59,32 %	452	29,06 %
Pagado	61,38 %	115	26,88 %
Subvencionado	58,80 %	243	29,59 %
Municipal	57,60 %	94	29,28 %

Tabla 4.21: Exploración Cubo Rendimiento por cohorte

Cohorte	Rendimiento	Alumnos	SD.Rendimiento
All cohortes	59,32%	452	29,06%
2004	69,63%	67	31,70%
2005	63,93%	77	28,56%
2007	61,25%	73	28,32%
2006	52,16%	73	28,69%
2008	44,87%	80	27,73%
2009	37,97%	82	24,36%

Tabla 4.22: Exploración Cubo Rendimiento por región

Region	Rendimiento	Alumnos	SD.Rendimiento
All region	59,32%	452	29,06%
IX	80,00%	3	47,21%
X	76,12%	14	22,12%
XII	74,81%	5	34,67%
VIII	69,57%	5	26,80%
IV	69,08%	17	38,69%
II	60,25%	5	15,64%
VI	59,52%	51	30,66%
V	58,98%	242	28,25%
III	55,00%	6	29,57%
RM	52,66%	75	26,99%
VII	50,70%	12	27,34%
I	47,02%	13	26,34%
XV	23,81%	2	5,05%
XI	7,14%	2	10,10%

Tabla 4.23: Exploración Cubo Rendimiento por tipo de curso

Curso	Rendimiento	Alumnos
All cursos	59,32%	452
DER	100,00%	35
COM	82,89%	67
ICA	79,36%	121
EII	76,45%	137
EIE	75,63%	91
ICI	72,06%	452
QUI	57,14%	393
FIS	54,87%	263
ICM	53,77%	220
MAT	41,38%	451

esta vista de análisis no aporta mayor información. Lo anterior sucede porque no existe data anterior a 2004 y por lo tanto los datos que se ven en la tabla 4.24 correspondientes a ese año son solo de las asignaturas de primer año de la cohorte del 2004. El rendimiento promedio se eleva porque los estudiantes de las cohortes anteriores avanzan en la malla y van aumentando cada año su tasa de aprobación.

4.3.4 Conclusiones del Análisis

Ya realizado el análisis explorando por cada dimensión cada cubo creado, se puede llegar a conclusiones que permitirán sentar la base para el proceso de minería de datos posterior. Para sacar conclusiones del análisis se han juntado los cubos y se han realizado cálculos de correlaciones entre las métricas. Se han fusionado los cubos de ingreso con los de rendimiento tanto del primer año como de todos los años. Por lo tanto, se analizó la correlación entre el rendimiento y el puntaje de la PSU y del rendimiento con las NEM.

4.3.4.1 Las Características de Ingreso y el Rendimiento del Primer Año

El objetivo de este análisis es sacar conclusiones entre rendimiento de los estudiantes en el primer año y las características propias del origen de los estudiantes. Para esto se realizó una descomposición por cada dimensión de la unión de los cubos «ingreso» y «primerano». Las dimensiones exploradas fueron el tipo de colegio, la región del colegio y el año de ingreso a la universidad (cohorte).

En relación al tipo de colegio, y lo descrito en las secciones 4.3.3.1 y 4.3.3.2, el tipo de colegio en promedio si puede representar el rendimiento del primer año. Si se clasifica por tipo de colegio la correlación entre el puntaje PSU y el rendimiento posterior es de un 0,9982. Las NEM se comportan de manera inversa con un *r* de -0,9894. Por lo tanto si un estudiante viene de un colegio particular y su PSU está en el extremo superior de la muestra y sus NEM en el inferior, es de esperar que obtenga un rendimiento superior. Lo opuesto pasaría con un estudiante de un colegio municipal.

Si el análisis se realiza en relación a la cohorte del estudiante la PSU presenta una correlación de un 0,0372 lo cual es débil. En cambio las NEM presentan una correlación mayor: 0,698 esta vez positiva. Por lo tanto si se separa en cohorte, y dado el anterior análisis (que la tendencia es a la baja), mientras si los estudiantes son de distintas cohortes sería mejor estimar por NEM que por PSU. Mientras más NEM tengan mayor sería el rendimiento en el primer año.

Tabla 4.24: Exploración Cubo Rendimiento por período

Periodo	Rendimiento	Alumnos
All periodos	59,32%	452
2004	52,75%	67
2005	53,88%	131
2006	53,65%	175
2007	59,67%	207
2008	63,83%	229
2009	63,25%	245

Si se analiza a través de la región del estudiante la PSU presenta una correlación de 0,41 con respecto al rendimiento posterior. Las NEM se correlacionan con este último con un coeficiente de 0,5354. Ambos tienen una correlación similar aunque el NEM se relaciona de mejor manera.

Dadas las correlaciones es posible deducir que las notas de la enseñanza media tienen una mayor correlación que el puntaje PSU con respecto al rendimiento del primer año. Esto estaría alineado con lo descrito en el marco referencial, donde se duda de la capacidad predictiva de la PSU. Aunque la correlación siempre es positiva, solamente en el caso del tipo de colegio es fuerte. Las NEM, por otra parte, presentan correlación positiva y mayor que la PSU en las dimensiones descritas, exceptuando la del tipo de colegio.

El tipo de colegio entonces, dado los resultados del análisis, presenta un buen indicador del rendimiento en el primer año. Lo es también la región de procedencia, aunque en algunas regiones la cantidad de estudiantes no sea la requerida para un mayor análisis. El análisis por cohorte en cambio no aporta información que pueda ser de apoyo para el modelo posterior, aunque se ve una tendencia a la baja del rendimiento a medida que las cohortes se acercan a la actualidad.

4.3.4.2 Las Características de Ingreso y el Rendimiento de la Carrera

El objetivo de este análisis es sacar conclusiones entre rendimiento de los estudiantes en la carrera completa y las características propias del origen de los estudiantes. Para esto se realizó una descomposición por cada dimensión de la unión de los cubos «ingreso» y «rendimiento». Las dimensiones exploradas fueron el tipo de colegio, la región del colegio y la cohorte del estudiante.

En relación al tipo de colegio, la correlación entre el rendimiento de la carrera y el puntaje PSU es de 0,9998. La correlación con las NEM alcanza el valor de -0,9944. Al igual que el análisis del primer año, si se analiza por tipo de colegio la PSU explicaría en una relación lineal directa el rendimiento de la carrera, mientras que las NEM una relación inversa directa.

Si el análisis se hace ahora por la cohorte del estudiante la relación entre el rendimiento y la PSU es de -0,0412 y la de las NEM es de un 0,9614. Por lo tanto mientras las cohortes que tienen un promedio NEM mayor a la media muestral tienen rendimiento superiores a los que están por debajo. La PSU presenta una relación inversa muy débil.

El análisis explorando la región del colegio de origen del estudiante arroja las siguientes relaciones. El rendimiento esta relacionado con la PSU en un 56,69%, mientras que las NEM en un 45,44%. Sin embargo, las regiones extremas siguen manteniendo su tendencia de bajo rendimiento, aunque la PSU es menor que el promedio y las NEM mas altas.

Por lo tanto, al igual que lo que pasa para el primer año el tipo de colegio es fundamental para explicar el rendimiento de toda la carrera del estudiante. La región también da un indicio pero no tan fuerte como el tipo de colegio. Por último la cohorte está influida por la permanencia de los estudiantes. Esto último dice relación con que los estudiantes que logran permanecer en la carrera comienzan a subir su rendimiento en los cursos superiores. Es por esto que la cohorte para primer año no es determinante, pero los años que el estudiante permanece en la universidad

si lo son.

4.4 Minería de Datos

En la presente subsección se detalla el trabajo realizado con respecto a la Minería de Datos. Se comienza detallando la preparación de los datos para la implementación del modelo predictivo, y las métricas de desempeño de estos modelos. Por último se detallan los resultados obtenidos con la aplicación de cada modelos. Como el objetivo de este proyecto es la predicción del desempeño académico de los estudiantes se hace vital definir que resultados esperamos predecir. Éstos pueden ser dos: se puede clasificar a los estudiantes por su estado de termino (aprobado o reprobado), o se puede realizar una regresión a la nota de cada curso inscrito. Es por lo anterior, que se realizarán ambas aproximaciones, aunque a priori se sabe que la regresión presenta una mayor dificultad de predicción.

Debido a la complejidad inherente del problema, y los variados factores que afectan el desempeño académico (que no pueden siempre ser cuantificados o registrados), es que se ha preferido probar con diferentes tipos de máquinas para escoger la que dé mejores resultados, aunque eso signifique un gasto de recurso extra y un tiempo de entrenamiento mayor.

4.4.1 Preparación de los datos

Al realizar un modelo predictivo se deben preparar los datos de entrada, ya que las máquinas de aprendizaje no tienen el mismo lenguaje que los humanos. Además debe separarse la entrada en dos grandes conjuntos; el primero para entrenar la red, llamado conjunto de entrenamiento, y el segundo para validar el modelo, llamado validación o reserva.

4.4.1.1 Variables de entrada

Los datos de entrada de las máquinas deben ser modificados para que el entrenamiento tenga significancia. No es necesario transformar las variables cualitativas, ya que cada valor de la variable será un nodo de entrada de la red. Las variables cuantitativas deben ser cambiadas de escala, y como se vio en la subsección 4.3.1 los datos presentaban distribuciones parecidas a la normal. Es por lo anterior que se ha decidido normalizar las variables cuantitativas de entrada de la red neuronal utilizando min-max entre 0 y 1, descrito en la ecuación 4.3.5. Además, para mejorar los resultados, se optó por transformar las variables nominales de escala en numéricas y posteriormente normalizarlas mediante la misma técnica.

4.4.1.2 Selección de atributos

Para este proyecto se tiene una gran cantidad de variables de entrada. Cuando se construye una Máquina de Aprendizaje con muchas variables de entrada, éstas tienden a entregar resultados menores a que si se entrenaran con las variables que más aportan en la realidad al modelo. Es por ésto que han nacido técnicas de selección de atributos, que permiten escoger un subconjunto de variables de entrada que sean las más representativas para el modelo y la instancia en específico.

Se realizó una selección de variables de entrada mediante dos técnicas: Information Gain y Evaluación de Atributo Chi-Cuadrado. Para ambas técnicas el resultado fue el mismo subconjunto de atributos. Se logró rebajar desde 24 a 11 atributos, incluso subiendo la efectivi-

dad de clasificación de todos los modelos probados. Este subconjunto de datos fue el que se utilizará desde aquí en adelante para el entrenamiento y validación de las distintas máquinas. Los atributos seleccionados, en orden de importancia, son: NombreAsignatura, CodAsignatura, SiglaAsignatura, PSU_Matematica, NumeroAsignatura, Area, AnoIngreso, AnoCurso, CreditosAsignatura, Promedio_PSU, PSU_Lenguaje.

4.4.1.3 Conjuntos de Entrenamiento y Validación

En la literatura aparecen distintas formas de dividir el set de datos para entrenamiento, pruebas y validación. A priori se podría pensar que dividir la data en tres conjuntos de igual tamaño podría ser una opción factible, pero estudios han demostrado que es la peor opción para datasets pequeños[11], el cual es el caso de este proyecto. El mismo estudio referenciado recomienda dejar un tercio del total para validación e ir variando su tamaño con distintas pruebas. Se realizaron tests con distintos volúmenes de datos y finalmente se utilizó un 80% para entrenamiento y un 20% para validación, ya que presentó el mejor resultado.

4.4.2 Resultados Obtenidos en la Clasificación

A continuación se presentan los resultados de la clasificación de estudiantes, tanto para las clases aprobación y reprobación. Primeramente se nombrarán las métricas para verificar el modelo predictivo seguido por los resultados de los diferentes tipos de Máquinas de Aprendizaje utilizadas. Para esta parte del trabajo se utilizó el software WEKA de la Universidad de Waikato.

4.4.2.1 Métricas

Las métricas utilizadas para medir los resultados de la clasificación son la efectividad o medida F1, la exactitud y el área de la curva ROC. A continuación se detallan cada una de ellas.

Efectividad(F1): Se define como la capacidad del clasificador de obtener los resultados. Se mide con dos componentes: precisión(π) y cobertura (ρ). A estas dos ultimas componentes se les denomina medidas F_1 (ver ec. 4.4.8), que es un caso especial de la medida F_β, cuando β=1.

Cuando la máquina clasifica correctamente un valor positivo se le denomina Verdadero Positivo(VP), por otra parte cuando correctamente clasifica un resultado negativo se le llama Verdadero Negativo(VN). Cuando clasifica incorrectamente un valor positivo se le denomina Falso Positivo (FN), y por último, cuando se clasifica incorrectamente un valor negativo se le un Falso Negativo(FP). Para una mejor descripción puede verse la tabla 4.25.

Precisión: La precisión se define, como puede verse en la ecuación 4.4.6,la proporción de Ver-

Tabla 4.25: Descripción de los tipos de clasificación

		Valor pronosticado	
		P	N
Valor real	P	Verdadero Positivo	Falso Negativo
	N	Falso Positivo	Verdadero Negativo

daderos Positivos entre los clasificados como positivos (es decir la suma de VP y FP)

$$\pi = \frac{VP}{VP+FP} \tag{4.4.6}$$

Cobertura: Consiste en la proporción de VP con respecto a los que en realidad eran positivos (los VP y FN). También es llamado Sensibilidad o Ratio de Verdaderos Positivos

$$\rho = \frac{VP}{VP+FN} \tag{4.4.7}$$

$$F_1 = \frac{2\pi\rho}{\pi+\rho} \tag{4.4.8}$$

Exactitud: Representa la proporción de elementos clasificados correctamente con respecto al total de datos. es decir: $Exactitud = \frac{VP+VN}{VP+VN+FP+FN}$.

AUC ROC: Consiste en el área bajo la curva ROC, que describe la relación entre el Ratio de Verdaderos Positivos (Sensibilidad) y el Ratio de Falsos Positivos (1 menos el Ratio de Verdaderos Negativos o Especificidad). Es una medida ampliamente utilizada para comparar modelos y medir la efectividad de los modelos creados.

4.4.2.2 Máquina de Soporte Vectorial

Se construyó una Máquina de Soporte Vectorial (SVM) del tipo lineal L2-loss para clasificar a los estudiantes. Utilizando esta máquina se logró el siguiente mejor resultado:

Tabla 4.26: Resumen de la validación de la clasificación con SVM

Instancias correctamente clasificadas(Exactitud)	60.4824 %
Instancias incorrectamente clasificadas	39.5176 %
Error absoluto medio	0.3952
Raíz del error cuadrático medio	0.6286

Tabla 4.27: Detalle de la exactitud de la clasificación con SVM

Clase	Medida F1	Precisión	Cobertura	Area ROC	Ratio VP	Ratio VN
A	0.441	0.596	0.35	0.58	0.35	0.191
R	0.694	0.608	0.809	0.58	0.809	0.65
Media Pond.	0.582	0.603	0.605	0.58	0.605	0.445

4.4.2.3 PART

PART es una lista de decisión. Utiliza la técnica de separar y conquistar. Construye un árbol de decisión C4.5 parcial en cada iteración haciendo una regla por cada mejor hoja[12]. Los resultados obtenidos utilizando PART se muestran a continuación.

Tabla 4.28: Resumen de la validación de la clasificación con PART

Instancias correctamente clasificadas(Exactitud)	63.6364 %
Instancias incorrectamente clasificadas	36.3636 %
Error absoluto medio	0.4284
Raíz del error cuadrático medio	0.4941

Tabla 4.29: Detalle de la exactitud de la clasificación con PART

Clase	Medida F1	Precisión	Cobertura	Area ROC	Ratio VP	Ratio VN
A	0.536	0.621	0.471	0.649	0.471	0.231
R	0.701	0.644	0.769	0.649	0.769	0.529
Media Pond.	0.627	0.634	0.636	0.649	0.636	0.396

4.4.2.4 Red Neuronal con Perceptrón Multicapa

Se construyó también una red con Perceptrón Multicapa(ANN-MLP) para comparar con los demás modelos. Los resultados se muestran a continuación:

Tabla 4.30: Resumen de la validación de la clasificación con ANN-MLP

Instancias correctamente clasificadas(Exactitud)	64.0074 %
Instancias incorrectamente clasificadas	35.9926 %
Error absoluto medio	0.4209
Raíz del error cuadrático medio	0.4866

Tabla 4.31: Detalle de la exactitud de la clasificación con ANN-MLP

Clase	Medida F1	Precisión	Cobertura	Area ROC	Ratio VP	Ratio VN
A	0,565	0,612	0,525	0,665	0,525	0,268
R	0,693	0,658	0,732	0,665	0,732	0,475
Media Pond.	0,636	0,637	0,640	0,665	0,640	0,383

4.4.3 Resultados Obtenidos en la Regresión

Para realizar una regresión a la nota estandarizada de los estudiantes, esto es la nota relativa al curso específico, se utilizó el software IBM SPSS Statistics. Se efectuó una regresión lineal automatizada, la cual se detalla a continuación.

El valor R^2 es de un 8,5 %. Esta correlación fue conseguida transformando los datos de entrada y seleccionando las variables más importantes. En la figura 4.8 puede verse la importancia de las variables tales como la PSU_Matematica con un 40 % por ejemplo.

Posteriormente se calcula mediante la curva de regresión los valores pronosticados y se comparan con los valores reales u observados. La relación de estas dos variables está graficada en la figura 4.9.

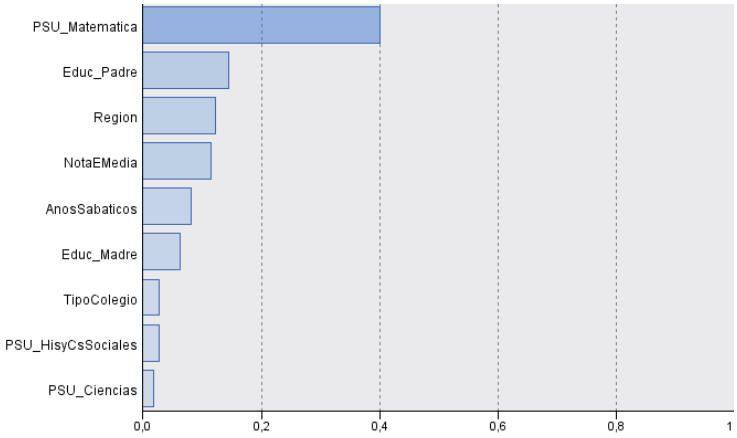

Figura 4.8: Importancia de las variables para la regresión

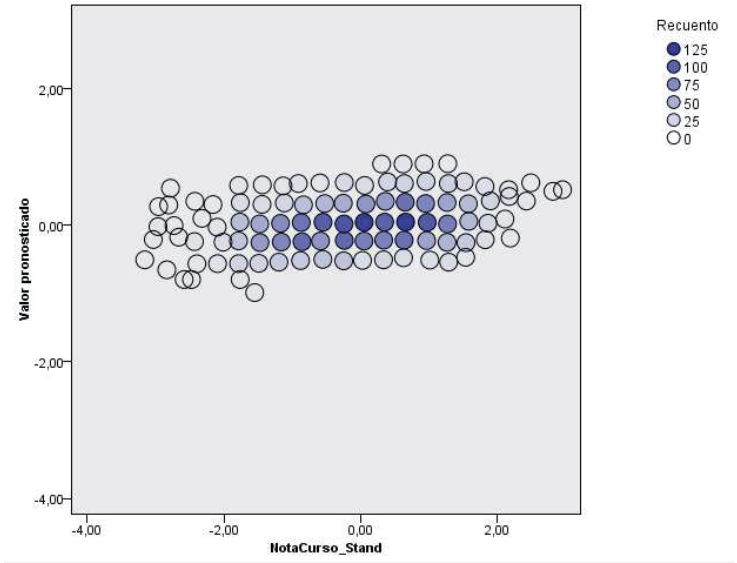

Figura 4.9: Valor pronosticado v/s valor observado

5 Conclusiones

Tal como se explicó en el comienzo de este trabajo, se partió del supuesto de que la información histórica que poseía la universidad permitiría realizar un modelo que permitiera la predicción del desempeño estudiantil de la Escuela de Informática de la PUCV. Para ésto se confeccionó un listado de variables que se creían eran de importancia para este modelo y explicarían la realidad a través de patrones de comportamiento. Al tener los datos se cargaron a un almacén de datos para permitir visualizarlos a través de dimensiones o cubos, después de un proceso de ETL.

Mediante la premisa de que la batería de la PSU no era el único predictor del rendimiento estudiantil posterior, esto basado en la literatura revisada, se realizó un análisis descriptivo en detalle. Este análisis permitió conocer la distribución de los datos de cada variable y la posible ocurrencia de datos erróneos que hayan sido obviados en el proceso de ETL. A continuación se procedió con la construcción de los modelos predictivos. Se tomaron dos caminos alternativos, el primero es predecir el rendimiento de los estudiantes mediante la finalización de la asignatura, esto es aprobación o reprobación. El segundo en tanto es una regresión a la nota estandarizada, esto es, relativa al curso que se esté tomando.

Para la clasificación se probaron distintas máquinas de aprendizaje, tales como las Máquinas de Soporte Vectorial(SVM) y las Redes Neuronales Artificiales (ANN), además de una regla de decisión basada en árboles C4.5 llamada PART. La SVM creada tuvo una exactitud del 60.48% y una efectividad ponderada del 58.2%. PART en tanto mostró una exactitud del 63.64% y una efectividad F1 ponderada del 62.7%. Por otra parte la ANN presentó una exactitud del 64.01% y una efectividad del 63.6%. Las curvas ROC tuvieron áreas de 0.58, 0.649 y 0,665 respectivamente. Para la regresión, en tanto, se obtuvo un valor R^2 del 8,5%, lo que era esperado ya que la regresión presenta una mayor dificultad de precisión que la clasificación.

Los resultados demuestran que para la clasificación la ANN se comporta de mejor manera en forma general al clasificar mejor en promedio a los estudiantes. PART no está lejos e incluso clasifica mejor que ANN a los estudiantes que aprobarían la asignatura, pero en desmedro de clasificar bien a los reprobados. SVM, por otra parte, queda por debajo de ambas máquinas anteriormente mencionadas, descartándose por el momento su uso. Para la regresión, en tanto, el modelo no se ajusta a la realidad y su capacidad predictiva no es importante.

Debido a la naturaleza humana, y las diferentes circunstancias del proceso educativo que no pueden ser medidas con variables cuantitativas, ademas de las variables no incluidas en el inicio del trabajo; es que debido a los resultados obtenidos, no es posible utilizar estos modelos construidos para poder predecir, con una confianza alta, el desempeño estudiantil. En el mejor caso, de 10 estudiantes, sólo se estaría prediciendo correctamente el rendimiento de alrededor de 6, lo cual no permite tomar este modelo como la única aproximación para la toma de decisiones. Es posible que el modelo propuesto en el presente trabajo permita tener una primera alerta sobre algún estudiante con problemas, pero no debe tomarse como un indicador eficaz para la toma de decisiones. Finalmente, los objetivos propuestos se han cumplido de forma cabal, por lo que se da por completado este proyecto.

Como trabajo futuro queda la predicción del desempeño estudiantil para las asignaturas de segundo año. Así como también un cambio en las variables de entrada, permitiendo incluir variables sicológicas, de entorno y de motivación que podrían influir con mayor propiedad en el desempeño estudiantil de los dos primeros años de la carrera. También es posible construir un Cuadro de Mando Integral, para ver como las decisiones tomadas con ayuda de estos modelos influyen en los indicadores clave de desempeño (KPI) educacionales de la Escuela de Ingeniería Informática de la PUCV.

Al finalizar el presente trabajo, y culminar esta parte tan importante de mi vida, quisiera expresar toda mi gratitud a la gente que participó activa o pasivamente de este proyecto. Aunque este proyecto sea sólo una parte de todo el trabajo que significó el paso por esta gran Universidad, también hay que agradecer a todas las personas que compartieron, apoyaron e hicieron de que estos años fueran los mejores de mi vida. Muchas gracias.

Referencias

[1] K. Mckenzie and R. Schweitzer. Who succeeds at university? factors predicting academic performance in first year australian university students. *Higher Education Research & Development*, 20:1+, 2001.

[2] Colin Power. *Success in higher education*. Canberra : Australian Government Publishing Service, 1987.

[3] G. Villapalos. El futuro de la universidad. *Política y reforma universitaria*, 1:333–340, 1998.

[4] M. Díaz. *Evaluación del Rendimiento en la Enseñanza Superior. Comparación de Resultados entre Alumnos procedentes de la LOGSE y del COU*. MINISTERIO DE EDUCACIÓN, CULTURA Y DEPORTE SECRETARÍA GENERAL DE EDUCACIÓN Y FORMACIÓN PROFESIONAL Centro de Investigación y Documentación Educativa (C.I.D.E.), 2001.

[5] D. Contreras, S Gallegos, and F. Meneses. *Determinantes de Desempeño Universitario*. Consejo Superior de Educación, 2009.

[6] J. G. Mora. *Universidad y trabajo*. Luxán, J. M., 1998.

[7] M. Berthold and D. Hand. *Intelligent Data Analysis*. Spinger Verlag, 2003.

[8] Fionn Murtagh. Origins of modern data analysis linked to the beginnings and early development of computer science and information engineering. *Electronic Journal for History of Probability and Statisics*, 4(2):1+, dic. 2008.

[9] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From datamining to knowledge discovery: An overview. *Advances in Knowledge Discovery and Data Mining*, pages 1–34, 1996.

[10] Roland Bouman and Jos van Dongen. *Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL*. Wiley, 2009.

[11] Patricia S. Crowther and Robert J. Cox. A method for optimal division of data sets for use in neural networks. *KES 2005*, LNAI 3684:1–7, 2005.

[12] Eibe Frank and Ian H. Witten. Generating accurate rule sets without global optimization. *Fifteenth International Conference on Machine Learning*, pages 144–151, 1998.

[13] DEMRE. *Nociones Básicas de Estadística Utilizadas en Educación*. Universidad de Chile, 2008.

[14] M. M. Duarte and Galaz J. Predictores del desempeño académico en el primer año de universidad en una institución pública estatal. In *6º Congreso Internacional Retos y Expectativas de la Universidad: El Papel de la Universidad en la Transformación de la Sociedad*, 2006.

[15] R. Duda. *Pattern Classification*. Wiley-Interscience, 2000.

[16] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Boston: Academic Press, 1990.

[17] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

[18] Sebastián Prado. Estudio de validez predictiva de la psu y comparacion con el sistema paa. Master's thesis, Universidad de Chile, 2008.

[19] G. Retamales and P. Olivares. Estudio de la confiabilidad de las pruebas de selección universitaria. Technical report, Unidad de estudios e investigación. DEMRE, 2009.

[20] B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 2000.

[21] Antonio Rúa. Búsqueda de patrones de rendimiento académico mediante técnicas de análisis multivariante. aplicación a 1º e4. Technical report, Universidad Pontificia Comillas de Madrid, 2001.

[22] Guiselle Vargas. Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación 31*, 1:43–63, 2007.

[23] W. Webster, R. Mendro, and T. Almaguer. Effectiveness indices: A «valueadded» approach to measuring school effects. *Studies in Educational Evaluation*, 20:113–145, 1994.