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

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

RICARDO ALFONSO BOCAZ LEÓN

INFORME FINAL DE PROYECTO 1
PROFESOR GUÍA: RODRIGO ALFARO ARANCIBIA
PROFESOR CORREFERENTE: CRISTIAN ALEXANDRU RUSU

SEPTIEMBRE, 2011

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. Anticiparse a los hechos puede ser aún mejor, ya que permitiría la toma de decisiones en forma oportuna por parte de la dirección de las escuelas. 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 toma de decisiones a los directivos.

En este proyecto se desarrollará la herramienta antes mencionada mediante la utilización de máquinas de aprendizaje. Para lograr ésto se debe realizar un proceso previo de análisis de datos apoyado por herramientas informáticas para disminuir los tiempos y aumentar la precisión. Este análisis de datos permite la identificación de características que influyen en el desempeño estudiantil universitario y que sientan las bases para el modelo predictivo.

Los factores característicos que influyen en el desempeño estudiantil, que se han encontrado hasta el momento, son el tipo de colegio de egreso, la región a la que pertenece el establecimiento donde egresó el estudiante, el área de estudio a la que pertenece la asignatura inscrita y el tiempo de permanencia del estudiante en la universidad. Estos factores permitirán comenzar con la siguiente fase del proyecto: la minería de datos.

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

1 Introducción

En las últimas décadas, la cantidad de estudiantes que siguen estudios universitarios ha ido en aumento[1]. Este incremento de la población universitaria va acompañado también por un crecimiento en la diversidad de las características de los alumnos. Estudiantes de diferentes sectores socioeconómicos, regiones y realidades tienen también necesidades y potencial académico distintos[1]. El reto de las universidades entonces es reconocer esta heterogeneidad y afrontarla como mejor puedan. Baker (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, por ende, 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 también han incidido en los progresos alcanzados por los estudios enfocados en el desempeño académico en educación superior. Por lo tanto aumentar el desempeño de los estudiantes es el objetivo principal de cualquier institución educativa. 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.

Al disponer la PUCV de un registro académico de las asignaturas dictadas en las carreras de la Escuela de Ingeniería Informática y los estudiantes que las han inscrito, permite tener una base de datos con información histórica propicia para realizar un proceso de análisis de datos. Este proceso se apoyaría en las nuevas tecnologías para obtener de manera rápida y precisa conclusiones que permitan el desarrollo de máquinas de aprendizaje. Estas herramientas de la minería de datos permitirían adelantarse a los hechos apoyando la toma de decisiones por parte de la dirección de la Escuela para mejorar sus procesos docentes y asignar los recursos de una manera eficiente.

2 Objetivos de la investigación

2.1 Objetivo General

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

[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

2.2 Objetivos Específicos

- Realizar un estudio acabado de los conceptos y las investigaciones que se relacionan con el presente proyecto.
- Aplicar 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 Programación

Las actividades descritas en la figura 2.1 se han planificado realizar en los plazos especificados en la misma. Estas fechas son tentativas y pueden ser supuestas a modificaciones, tanto por el investigador o sus profesores.

3 Marco Teórico

En este capitulo 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. Todas las consideraciones antes descritas están relacionadas entre sí. Con la data histórica guardada por la PUCV se realiza un proceso de Extracción, Transformación y Carga (ETL). Éste 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 ya se encuentran en la universidad, como también de los que próximamente se integrarán.

	Nombre de tarea	Duración	Comienzo	Fin	Pred
1	Definir los objetivos del proyecto	14 días	vie 11/03/11	jue 24/03/11	
2	Estudiar el marco conceptual	14 días	vie 25/03/11	jue 07/04/11	1
3	Confeccionar y revisar Informe de Avance	14 días	vie 08/04/11	jue 21/04/11	1;2
4	Entregar Informe de Avance	0 días	jue 21/04/11	jue 21/04/11	3
5	**Obtener los datos necesarios**	**24 días**	**lun 09/05/11**	**mié 01/06/11**	**4**
6	Definir datos necesarios	7 días	lun 09/05/11	dom 15/05/11	
7	Solicitar y recibir los datos	17 días	lun 16/05/11	mié 01/06/11	6
8	Realizar proceso de ETL	15 días	jue 02/06/11	jue 16/06/11	5
9	Realizar analisis ROLAP	15 días	vie 17/06/11	vie 01/07/11	8
10	Confeccionar y revisar Informe Final	14 días	sáb 18/06/11	vie 01/07/11	
11	Entregar Informe Final	0 días	vie 01/07/11	vie 01/07/11	10

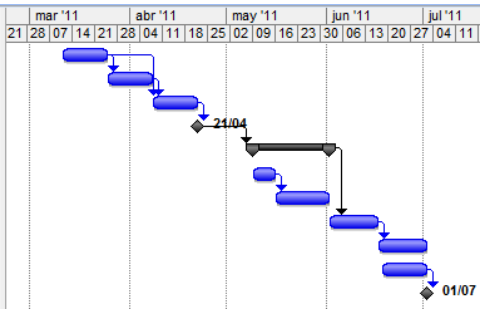

Figura 2.1: Carta Gantt del proyecto

3.1 Marco Conceptual

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 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 variadas formas a cada estudiante en particular. Existen factores que son propios de cada estudiante, como el conocimiento acabado de la carrera que estudia, o la motivación por el área de estudio que realiza, como también factores externos: 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. La manera de medir el desempeño estudiantil es a través de la construcción de indicadores de desempeño, que dan 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 aplican a cualquier tipo de investigación.

3.1.2 Análisis Inteligente de Datos

El análisis inteligente de datos (AIDA) es una metodología que proviene de la complementación de diferentes disciplinas, tales como la Estadística y las Máquinas de Aprendizaje[4]. La Estadística aporta el modelado de la realidad, la recolección 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, ademas 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.

3.1.2.1 Análisis de Datos

El análisis de datos se define como «El proceso de computar diversos resúmenes y valores derivados de la colección de datos dada»[4]. 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. Se pueden distinguir dos tipos de análisis, los descriptivos y los inferenciales. Los descriptivos

apuntan a rescatar información propia de la muestra que se analiza. En cambio los inferenciales tratan de explicar como se comporta la población basado en una muestra representativa de ella.

3.1.2.2 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 son la extracción, la transformación y la carga de los datos (ETL).

3.1.2.3 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. La integración de datos consiste en la extracción, la transformación y la carga.

La extracción cosiste en 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. La transformación es cambiar la forma o representación de los datos para hacerlos coincidir con la forma del almacén de datos que se desea crear. Por último la carga es la etapa donde se escriben los datos en el almacén creado. Existen actividades de apoyo al proceso de extracción, las cuales son: captura de datos modificados y puesta en escena. Para la etapa de transformación también existen actividades que apoyan esta labor, algunas de ellas son la validación, limpieza, decodificación, agregación e identificación. En la etapa de carga se encuentran las siguientes actividades que apoyan el proceso: carga de tablas de hechos y carga de tablas de dimensión.

3.1.2.4 Procesamiento Analítico En Línea (OLAP)

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.

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. Se debe definir qué esquema seguirá el almacén de datos. Existen esquemas de amplia utilización tales como el estrella o el copo de nieve.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.

3.1.3 Minería de Datos

El término minería de datos 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 de identificar patrones válidos, novedosos, potencialmente útiles y que en ultima instancia sean entendibles»[5] Por lo tanto, basándose en esta definición 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 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 agrupamiento y regresión[6].

La clasificación es el proceso de separar o dividir el conjunto de datos en grupos llamados clases. 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. 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.

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. 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. El entrenamiento y las pruebas 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.

3.2 Marco Referencial

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 aporten los instrumentos de la batería de selección actualmente en uso.

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.

El objetivo de esta investigación era verificar si existía una estructura subyacente que explicara el rendimiento de los estudiantes. Esta estructura ha podido ser explicada a través de 4 factores comunes (Factor Cuantitativo, Factor Lingüístico Humanístico, Factor Empresarial y Factor 2º Idioma), cada uno de los cuales incide en las cuatro vertientes que cabe destacar en 1º curso de E4: vertiente cuantitativa, Humanística, Empresarial e Idioma. Mediante un análisis de conglomerados se han encontrado ocho tipologías académicas, es decir, los 65 alumnos se ha repartido en 8 conglomerados con características netamente diferenciadas. Las tipologías encontradas se pueden corresponder con los siguientes patrones de comportamiento: buen rendimiento académico, situaciones atípicas, rendimientos académicos pésimos, rendimientos académicos malos, rendimientos académicos que destacan en alguna faceta y rendimiento académico normal.

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 (PUC), 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. 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.

4 Desarrollo del proyecto

4.1 Datos Necesarios

Los datos que son necesarios para la investigación se describen en este apartado. Se han dividido en categorías o dimensiones, dependiendo del origen de éstos, los cuales pueden ser de procedencia, de la batería de selección o del propio historial universitario. Para cada estudiante debe generarse un identificador, 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.

Los datos necesarios en la dimensión procedencia son: clasificación socioeconómica, establecimiento de egreso, tipo de establecimiento, ranking de egreso, región del establecimiento, año de término de la educación media, modalidad estudios del estudiante, notas de la enseñanza media (NEM).Los de la dimensión batería de selección son: puntaje de la PSU/PAA, puntaje de las pruebas en específico, número de preferencia. Mientras que en la dimensión universitaria son necesarios los siguientes: año de ingreso, asignaturas cursadas, calificaciones de las asignaturas cursadas, año de egreso, año de titulación, año de congelamiento o deserción.

4.2 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. Consiste en un proceso cíclico que hasta el momento de la elaboración del presente documento no ha terminado y no lo hará hasta concluir el proyecto. La herramienta utilizada de apoyo al proceso de integración de datos es Pentaho Data Integration (PDI).

4.2.1 Proceso de Extracción

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 universidad. Este origen consta de una hoja de cálculo incluyendo los distintos datos ya extraídos de las bases de datos de la universidad. 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 se realizará el proceso de transformación. La primera fase de la extracción se realiza separando la data de los estudiantes dependiendo de la carrera a la que están adscritos. La segunda etapa de extracción se hace desde la misma base de datos para el proceso de transformación.

4.2.2 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.

En la etapa de validación 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). 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). Al realizar la limpieza se hizo 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. En la 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. Y por último en la identificación se generó, para cada asignatura cursada por cada estudiante, un código identificador autoincremental para la tabla de hechos del almacén de datos. También para cada asignatura dictada en momentos distintos se realizó lo mismo.

4.2.3 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. 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.

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.

4.3 Almacén de Datos

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

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.

4.4 Análisis de los datos

4.4.1 Métricas de desempeño utilizadas

Debido a la falta de datos en el período de desarrollo de éste proyecto, la única medida de desempeño utilizada fue la Tasa de Rendimiento Porcentual (TRP). La TRP se define como la razón de las asignaturas aprobadas con respecto a las inscritas y su fórmula matemática puede verse en la ecuación 4.4.1.

$$TRP = \frac{AsignaturasAprobadas}{AsignaturasInscritas} * 100 \tag{4.4.1}$$

4.4.2 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. Las correlaciones han sido medidas por el Coeficiente de Correlación de Pearson detallado en la ec. 4.4.2 . 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.

$$r_{xy} = \frac{\sum x_i y_i - n\overline{xy}}{n s_x s_y} \tag{4.4.2}$$

4.4.2.1 Las características de ingreso y el rendimiento del primer año

El tipo de colegio, 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.

4.4.2.2 Las características de ingreso y el rendimiento de la carrera

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.

5 Conclusiones

Para lograr una culminación exitosa de esta etapa del proyecto era necesario realizar una definición de los principales conceptos involucrados. Esto se describió en la sección 3, tratando tanto el desempeño académico, el análisis inteligente de datos y la minería de datos. Esto sentó las bases del trabajo de esta fase del trabajo, ya que permitió el desarrollo de éste siguiendo las metodologías de los conceptos estudiados. También se describieron algunos trabajos relacionados con la temática de esta investigación.

En el desarrollo de esta primera fase se logró definir, primeramente, los datos necesarios para la elaboración de esta etapa. Después de la definición de los datos se siguió la metodología del análisis inteligente de datos. El análisis de los datos fue realizado según lo estudiado en el marco teórico. Después de realizado todo el proceso del análisis se detallaron algunas conclusiones de éste.

Las conclusiones del análisis permiten sentar las bases para la siguiente fase del proyecto: la minería de datos. Se pudo identificar cuatro características que podrían influir en el rendimiento, tanto del primer año de la carrera como en su totalidad. Pudo establecerse que el puntaje PSU no es el único factor ponderante en el desempeño posterior y que la institución debe plantearse cambiar los métodos de selección actuales.

Las principales características de los estudiantes que pudieron ser identificadas en esta fase son el tipo de colegio de egreso de la enseñanza media, la región a la que pertenecen y el tiempo de permanencia en la universidad. Se pudo ver que los estudiantes que provienen de colegios particulares superan en rendimiento a los de los subvencionados, y éstos a los de los municipales. También se encontró que los estudiantes de las regiones extremas presentan rendimientos inferiores. Por último mientras más años se permanezca en la universidad el rendimiento va aumentando; las asignaturas de primer año presentan bajas tasas de aprobación mientras que las de últimos años las más altas. El área de estudio también influye; asignaturas de derecho, comercial e industrial están en el grupo superior de aprobación, en un segundo grupo están las asignaturas propias de la informática y en el último grupo la de las ciencias básicas, con las matemáticas en el fondo de la clasificación.

La próxima etapa consistirá en el desarrollo de modelos predictivos que permitan pronosticar el desempeño de un estudiante a través de su carrera. Deberá probarse tantas herramientas como sea posible y comparar su capacidad predictiva. Las conclusiones de el presente informe deben ser la base para lograr la culminación exitosa de la próxima fase de este proyecto.

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. Berthold and D. Hand. *Intelligent Data Analysis*. Spinger Verlag, 2003.

[5] 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.

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