Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería

Sergio Celis [*]
Luis Moreno [*]
Patricio Poblete [*]
Javier Villanueva [*]
Richard Weber [*]

Resumen

En la última década el avance de los sistemas de gestión docente y sistematización de datos en educación superior han motivado el uso de herramientas de la minería de datos para entender procesos de aprendizaje y los contextos en los cuales estos ocurren. En el mundo anglosajón, comunidades en torno al *learning analytics* o el *educational data mining* han surgido para desarrollar áreas de investigación e intervención en educación superior. En estas comunidades, un área de particular interés es la generación de modelos predictivos de deserción y rendimiento académico que permitan intervenciones de apoyo temprano a los estudiantes. En este artículo hacemos uso de herramientas de *learning analytics* para construir un modelo que predice la caída en causal de eliminación, por motivos académicos, en estudiantes de primer año del Plan Común de Ingeniería y Ciencias de la Universidad de Chile. El modelo clasifica correctamente a más del 86 % de los casos, con niveles bajos de error tipo II, y una precisión de 38 %. Dado que se usa información hasta el inicio del segundo semestre, el modelo permite desarrollar intervenciones focalizadas en aquellos estudiantes en mayor riesgo.

Palabras Clave: Modelo predictivo, Rendimiento académico, Learning Analytics, Educational Data Mining.

[*]Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile.

1. Introdución

El enorme crecimiento en la disponibilidad de datos ha generado recientemente muchas oportunidades de aplicar métodos para el análisis de estos datos. El área de la educación no es una excepción. Analizando los datos que se genera entorno a la educación permite descubrir nuevas oportunidades para mejorar la gestión docente.

En este artículo describimos la aplicación de minería de datos aplicada a datos académicos de los estudiantes de ingeniería y ciencias de la Universidad de Chile y mostramos cómo la gestión docente puede anticipar - y posiblemente evitar - efectos negativos, como por ejemplo la doble reprobación de un curso que termina en la eliminación de la carrera y que es el fenómeno estudiado en el presente trabajo.

En la Sección 2 del artículo describimos el estado-del-arte del área de *learning analytics*. La Sección 3 describe la situación actual en una escuela de ingeniería. En la Sección 4 mostramos la construcción del modelo predictivo. Los resultados de la aplicación de nuestro modelo presentamos en la Sección 5. La Sección 6 concluye este trabajo y muestra posibles trabajos futuros.

2. Estado-del-Arte de Learning Analytics

A comienzos de este siglo, dos comunidades de investigación surgieron para usar herramientas matemáticas y computacionales para el análisis de datos educativos en educación superior: *Educational Data Mining* (EDM) y *Learning Analytics*. Ambas comunidades comparten el objetivo de usar la creciente recolección de datos en educación superior para mejorar los sistemas de evaluación, el entendimiento de los procesos educativos, y la priorización y diseño de intervenciones educativas [32]. Las diferencias entre ambas comunidades de investigación radica en los énfasis metodológicos y focos de investigación. En cuanto a metodología, mientras EDM privilegia el descubrimiento automatizado de patrones con poca intervención de juicio experto, *Learning Analytics* fortalece el juicio experto y testea hipótesis educacionales con ayuda de modelos de descubrimiento automático [4], tales como la selección de atributos [1]. Esto hace que el enfoque *Learning Analytics* sea más holístico y sistémico (p.ej. [27] que el enfoque basado en componentes individuales y la interacción entre ellos, característicos de la minería de datos [32]. En consecuencia, mode-

los generados por investigadores EDM son usualmente usados para desarrollar sistemas de tutoría inteligente, y los de *Learning Analytics* para apoyar la toma de decisiones de administrativos, profesores, y estudiantes. Sin embargo, ambas comunidades poseen límites porosos y múltiples convergencias entre ellas [32]. En lo que sigue, y sólo para efectos de este artículo, usaremos el concepto de *Learning Analytics* y lo entenderemos indistintamente a EDM.

Las tareas más comunes en *Learning Analytics* son clasificación, clustering, minería de textos, y visualización [22]. En cuanto a técnicas, las más usadas son árboles de decisión, redes neuronales, y redes Bayesianas [30]. Estás técnicas en contextos educacionales son frecuentemente complementadas con regresiones, correlaciones y otros técnicas estadísticas [30]. La principal fuente de datos para la investigación en *Learning Analytics* está en el uso de plataformas computacionales de aprendizaje, tales como sistemas de gestión de curso o CMS (*course management systems* en inglés) o sistemas de aprendizaje en línea [30], tales como los *Massive Open Online Courses* (MOOCs) [35]. Según Romero y Ventura [30] algunos de los problemas de investigación que concentran el interés de la comunidad de *Learning Analytics* son visualización de datos, retroalimentación a instructores, recomendaciones para estudiantes, predicciones de rendimiento de los estudiantes, modelos mentales de los estudiantes, y detección de comportamientos indeseados. En los últimos años su aplicación se ha extendido a otras áreas como el apoyo a metodologías activas basadas en problemas o proyectos [8], la toma de decisiones e intervenciones a nivel institucional (p.ej. [17], o el entendimiento de teorías del aprendizaje, tales como aprendizaje auto-regulado [29]).

Una contribución esencial del *Learning Analytics* a la línea clásica de teorías y modelos educacionales es que incorpora una nueva escala temporal a los procesos de aprendizaje. Si las teorías educacionales usan modelos invariantes en el tiempo o en largas etapas de desarrollo (en educación superior típicamente en semestres o años), las técnicas de minería de datos, son capaces de mostrar aprendizajes momento a momento [6]. Es decir, cambios en las capacidades de aprender, concentración y hasta estados de ánimo mientras el estudiante completa una evaluación en línea, trabaja en grupo, o interactúa con múltiples sistemas en el campus (p.ej. bibliotecas, unidades de tutoría académica). Más aún, esta información puede ser obtenida y procesada en tiempo real, permitiendo decisiones y acciones inmediatas o en el corto plazo [6]. De acuerdo a Berland et al. [8], *Learning Analytics* "permite una rigurosa, replicable, y precisa descripción del comportamiento de los estudiantes, así como también un análisis de cómo estos comportamientos interactúan con otros constructos de interés. El comportamiento de los estudiantes puede ser monitoreado en cuanto crece y cambia en el tiempo" (p. 211, traducción propia)

Otra emergente área de investigación en *Learning Analytics* es la combinación de datos institucionales con información proveniente del juicio humano. Esta combinación se realiza, por ejemplo, con instructores usando aplicaciones que evalúan el trabajo de los estudiantes en sala o talleres [6].

3. Retención y Rendimiento Académico en el Primer Año de Educación Superior

En las últimas décadas, la deserción en educación superior se ha transformado en un asunto prioritario de política educacional, tanto a nivel institucional como gubernamental. El impacto negativo de la deserción es relevante tanto porque los aranceles aumentan, como por el significado que socialmente ha adquirido la educación superior, entendida hoy como una instancia clave de desarrollo personal, social, económico y cultural. Se estima que en Chile la deserción al tercer año es cercana al 40 %, con una gran variabilidad según el tipo de institución (p.ej., universitaria, institutos profesionales, y centros de formación técnica) y áreas disciplinarias [31]. Por ejemplo, según Rolando et al. [28], en la cohorte del 2008 que ingresó a carreras profesionales, un 38 % desertó en el primer año en institutos profesionales, mientras que sólo un 14 % en universidades. De acuerdo al estudio Retención en Educación Superior con Perspectiva de Género [24] se evidencia que desde la cohorte 2007 hasta el año 2010 existió un aumento de la tasa de retención desde un 67 % a un 71 %, sin embargo, para el año 2013 ésta disminuyó a un 69 %. Según área disciplinar, la retención en los programas académicos en las áreas de tecnologías está entre las más bajas del país. En promedio, sólo un 65 % de los estudiantes permanece en sus programas luego del primer año [23].

La investigación de la persistencia y deserción tiene una larga historia en naciones con desarrollados sistemas de educación superior. Un gran número de estudios ha identificado los factores críticos que explican la persistencia y deserción. Parte importante de la complejidad a la que se ven enfrentados estos estudios, es la definición operacional de la deserción. Existen distintos tipos de deserción (p.ej., voluntaria o involuntaria; de transferencia o abandono), las cuales son registradas en diferentes tiempos (semanas, semestres, años), y que pueden ser transitorias o permanentes. Una discusión conceptual sobre las definiciones de la deserción en Chile puede ser consultada en [16].

Pascarella y Terenzini [26] revisaron más de tres décadas de este tipo de investigaciones, principalmente aquellas realizadas en Estados Unidos. Entre los resultados de su investigación, proponen un listado extenso de factores y

mecanismos que influyen en la persistencia y deserción, entre los que destacan características individuales de pre-ingreso y características institucionales. Las características individuales de pre-ingreso a la institución de educación superior tienen un consistente y estadísticamente significativo efecto en la persistencia. Al respecto, estudios previos han identificado la habilidad académica, status socioeconómico, grado de motivación, y expectativas de logro. Más aún, estas características académicas y sociales de los y las estudiantes tienen mayores efectos que las características institucionales en la persistencia y deserción. Otro factor importante es el ingreso retrasado a la educación superior, es decir, el tiempo que transcurre desde que el o la estudiante termina la educación secundaria, hasta que se matricula en alguna institución de educación superior. Estudios anteriores también muestran que deserciones previas tienen un efecto negativo en las chances de persistencia. Las características institucionales han recibido gran atención dado que pueden ser controladas por las instituciones y por políticas públicas. Aquellas características que han mostrado mayor impacto son la selectividad de las instituciones, incluso controlando por factores obvios como la habilidad académica de los y las estudiantes; su integración al campus y sus participaciones en actividades extracurriculares; actividades del primer año que introducen a estudiantes a la vida académica; becas para estudiantes de bajos ingresos; interacciones con profesores y profesoras fuera de la sala de clase; y la interacción entre pares. El rendimiento académico, las notas, es el mejor predictor de la persistencia, con un mayor efecto durante los dos primeros años de estudio. En relación a las diferencias disciplinarias, estudiantes en carreras de las ciencias, tecnologías, ingeniería y matemática (STEM en inglés), tienen una mayor tasa de deserción que estudiantes en otras disciplinas. Es importante mencionar que la mayoría de los factores ya discutidos interactúan con características sociodemográficas de los estudiantes, tales como etnicidad y sexo.

Algunos autores han propuesto constructos no observables para explicar la persistencia y deserción. Los modelos teóricos más influyentes son los modelos de deserción de Bean y el proceso de deserción de Tinto. Bean [7], basado en estudios de rotación organizacional, construye y testea un modelo de análisis de trayectorias causales para la deserción. En este modelo, la identificación de un o una estudiante con la institución, la certeza en la decisión de carrera, valores instrumentales (por ejemplo, creencia en que la educación es fundamental para conseguir un buen trabajo), y la intención de abandonar son factores que median los efectos de variables individuales, organizacionales y ambientales. Desde otra perspectiva teórica, los modelos de Tinto se basan en los estudios sobre suicidio de Durkheim y en los estudios de ritos de transición en sociedades tribales de Van Gennep. Tinto [34] extiende previos modelos de deserción,

proponiendo tres estados en la trayectoria de los y las estudiantes en educación superior: separación, transición, e incorporación, que son críticos en las decisiones de continuar o abandonar. Ambos, los modelos de Bean y Tinto, han sido consistentemente confirmados a través de estudios cuantitativos [9]. En la última década, investigadores han testeado estos modelos, analizando datos longitudinales y usando técnicas estadísticas más avanzadas. El estudio pionero de DesJardins, Ahlburg y McCall [15] basado en datos longitudinales de la University of Minnesota, arrojó que las variables definidas por estudios previos, tales como los descritos anteriormente, afectan la deserción, pero en magnitudes diferentes, según los años en la carrera. Por ejemplo, la locación de la residencia de origen tuvo un efecto significativo en la deserción en los tres primeros años de carrera, y la edad de ingreso sólo en los dos primeros. Numerosos estudios han continuado usando éstas y otras técnicas estadísticas para entender con mayor profundidad los fenómenos relacionados con la deserción (p.ej., [11, 12, 18, 20, 33]).

En Chile, también se han comenzado a testear estos modelos y a utilizar sofisticados métodos cuantitativos para entender la deserción en el sistema nacional de educación superior. Acuña [2] y Larroucau [19], basados en datos nacionales del sistema secundario y universitario chileno, confirman que el fenómeno de la deserción es multicausal y que las variables discutidas anteriormente tienen validez en el contexto local. Específicamente, Larroucau [19] encontró que en las características individuales de pre-ingreso, tales como el establecimiento de origen y el promedio de notas y ranking en la enseñanza media, eran mejores predictores de la deserción que el puntaje PSU. Mizala, Hernández, y Makovec [25] estiman la probabilidad de deserción en las carreras de pedagogía. Sus resultados confirman a la habilidad académica (medida por puntaje PSU) como uno de los factores más influyentes en la deserción, efecto que sería moderado por el quintil socioeconómico del estudiante. Díaz [13] y Celis [10] calcularon modelos de duración con datos de las carreras de ingeniería de la Universidad Católica de la Santísima Concepción y la Universidad de Chile, respectivamente. Ambos estudios muestran que el tipo de establecimiento de educación media impacta en la deserción. Celis [10] muestra que estudiantes provenientes de colegios particulares tienen menores tasas de deserción en los últimos años de la carrera que aquellos provenientes de la educación pública. Díaz [13] encontró que a mayor puntaje en la PSU y a mayor ingreso familiar, menores son las chances de deserción. El estudio mostrado en [31] usa la técnica de *propensity score matching* para estudiar el impacto de los créditos y becas en la persistencia. Los resultados sugieren una asociación positiva de los créditos y becas de excelencia en la persistencia. El Centro de Microdatos de la Universidad de Chile [14], mediante una

encuesta, determinó que las principales causas de deserción en el primer año universitario se deben a problemas vocacionales (p.ej., no quedar en la carrera de preferencia), situación económica familiar, y rendimiento académico. Recientemente, herramientas estadísticas tradicionales de la minería de datos también han comenzado a usarse para analizar la deserción y otras variables educacionales [21] (ver [3] para un caso aplicado en una universidad chilena).

En resumen sabemos que hay factores previos al ingreso, características individuales, y condiciones de vida y académicas que influeyen en la retención de primer año. Además sabemos que las carreras de ingeniería y ciencias tienen promedios altos de deserción en el primer año universitario. Sin embargo, muchas de estas investigaciones se han realizado en naciones con sistemas de educación superior desarrollados. Más investigación es necesaria para entender el fenómeno de la deserción en Chile, en especial en carreras de ciencia e ingeniería. Aquí es donde *Learning Analytics* brinda oportunidades no solo para entender empíricamente la deserción, sino que también para generar modelos predictivos que permitan generar alertas tempranas e intervenciones que le brinden apoyo oportuno a estudiantes en riesgo de deserción o de insuficientes desempeños académicos. A continuación se presenta un modelo predictivo desarrollado para detectar bajos rendimientos académicos en el primer año del Plan Común de las carreras de ingeniería y ciencias de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile.

4. Construcción del modelo predictivo

4.1. Situación actual en la Facultad de Ciencias Físicas y Matemáticas (FCFM)

La FCFM es una unidad académica altamente selectiva, con una alta producción científica y sofisticados sistemas de gestión docente relativo al contexto regional y latinoamericano. La población estudiantil es cercana a los 4.900 estudiantes de pregrado, seleccionados del 3 % superior de la enseñanza media de acuerdo al la Prueba Nacional de Selección Universitaria (PSU). La FCFM la componen además cerca de 1.200 estudiantes de postgrado y 220 profesores de jornada completa, de los cuales un 97 % posee un grado de doctor. La FCFM ofrece 9 carreras de ingeniería, geología y tres licenciaturas científicas. Todos los estudiantes de pregrado ingresan a un Plan Común de dos años de duración. Actualmente, el primer año está estructurado en dos semestres. En el primer semestre los estudiantes son asignados en siete secciones con similares

capacidades académicas según ranking de ingreso. Todos los estudiantes tienen los mismos ramos en el primer semestre: introducción al cálculo, introducción al álgebra, introducción a la física newtoniana, introducción a la ingeniería, química, y herramientas computacionales para ingeniería y ciencias. En total, la carga académica suma 30 SCT (Sistema de Créditos Transferibles), lo que equivale a 50 horas de trabajo semanal durante 15 semanas. En general, los estudiantes aprueban el 85 % de los cursos inscritos en primer año. En las últimas dos décadas, la FCFM ha venido realizando sostenidos esfuerzos para mejorar las tasas de retención y el rendimiento académico de los estudiantes. Por ejemplo, el 2007 se realizó un cambio curricular que implicó un giro hacia estrategias de enseñanza centradas en el estudiante, además de importantes mejoras de infraestructura y el lanzamiento de nuevas unidades de apoyo docente y al estudiante. Actualmente las tasas de retención de primer año son cercanas al 95 %. Pese a que este indicador es muy superior a las carreras de ingeniería y tecnología a nivel nacional (en torno al 65 %), la FCFM está empeñada en seguir mejorando esta tasa, consciente de la gran calidad académica de los estudiantes que recibe y de que el pequeño grupo que no persiste luego del primer año representa un desafío particular. El estudio aquí descrito se circunscribe en estos esfuerzos. Así, el objetivo de esta investigación es usar la información personal y académica disponible de los estudiantes para detectar estudiantes en riesgos de abandonar el plan de estudios. Para tales efectos se usó información histórica para generar y calibrar un modelo predictivo que permitiese la instalación de un sistema de alertas tempranas que le de soporte a los estudiantes que más lo necesiten. Para la construcción del modelo predictivo se usaron datos de las cohortes de ingreso 2010, 2011, 2012, 2013, y 2014. A continuación se presenta el modelo predictivo en sí, discutiendo la variable dependiente, las variables independientes consideradas, y la construcción del modelo.

4.2. Variable dependiente

En la primera fase del estudio se decidió acotar la variable dependiente a la doble reprobación de al menos un curso del primer semestre. Esta definición se justifica en dos ideas importantes. Primero, la doble reprobación de un curso es causal de eliminación de los estudiantes, que a la vez afecta negativamente las tasas de retención de primer año. Aunque un alumno en causal de eliminación puede elevar una solicitud especial para rendir un curso por tercera vez y proseguir en la Escuela, estas solicitudes requieren un esfuerzo no menor en la gestión docente y un porcentaje importante de estos alumnos igual termina eliminado de la Facultad. La segunda razón tiene un

argumento metodológico. La deserción en un lugar como la FCFM, así como en otras escuelas de ingeniería, es multidimensional y diversa. Tal como se indicó en la revisión de la literatura sobre deserción, las razones van desde lo económico (p.ej., falta de financiamiento) pasando por crisis vocacionales, situaciones excepcionales, hasta rendimiento académico. Así focalizarse en las causas académicas (las cuales no están necesariamente disociadas del resto), permite darle mayor precisión al modelo, al menos conceptualmente. La Tabla 1 muestra la distribución de la reprobación para las poblaciones estudiadas. Dado que reprobar al menos un ramo es condición necesaria para la reprobación de un ramo por segunda vez, la población de estudiantes considerada para esta investigación se reduce a entre 195 a 255 estudiantes por cohorte, que son los que reprobaron por lo menos un curso en su primer semestre.

Tabla 1: Reprobación y Doble Reprobación en Primer Año

Año Ingreso	Cohorte Ingreso [1]	Al menos 1 curso reprobado 1^{er} semestre	Doble reprobación 2^{do} semestre[2]
2010	687	195 (28 %)	43 (24 %)
2011	720	220 (31 %)	26 (14 %)
2012	704	213 (30 %)	41 (21 %)
2013	700	255 (36 %)	26 (11 %)
2014	762	216 (28 %)	27 (14 %)
Total	3.573	1.099 (31 %)	163 (17 %)

(1) Número de estudiantes que se mantuvieron activos durante el primer semestre.
(2) El porcentaje corresponde a estudiantes que reprobaron por segunda vez algún ramo de primer semestre sobre el total de estudiantes que reprobaron al menos un ramo de primer semestre y se mantuvieron activos durante el segundo.

4.3. Variables independientes

Las variables independientes (o atributos) consideradas fueron seleccionadas basado en la revisión de la literatura y la información disponible. Así las variables independientes se dividen en tres grupos: características individuales, variables de pre-ingreso y variables de rendimiento académico. En cuanto a características individuales sólo incluimos género, tiempo desde el egreso de la enseñanza media y región de procedencia. En variables de pre-ingreso usamos tipo de establecimiento de enseñanza media (i.e., particular, subvencionado, público emblemático y público no emblemático), experiencias previas en educación superior, puntajes en la PSU, vía de ingreso (i.e., PSU o ingresos

especiales), ranking y promedio de notas en la enseñanza media. Finalmente se construyeron otras once variables (continuas, ordinales y binarias) basadas en información detallada sobre las notas parciales de los estudiantes en los dos primeros semestres de la población objetivo. Dentro de aquellas variables podemos mencionar ratio de créditos aprobados versus reprobados, variación de notas del primer al segundo semestre tanto en ramos aprobados como reprobados, y diferencias con la nota mínima de aprobación, la cual en este caso es 4, dónde 1 es la mínima y 7 la máxima.

La decisión de cuánta información académica incluir en el modelo merece mayor discusión. En nuestro caso, la doble reprobación ocurre al final del segundo semestre del primer año. Mientras antes en el año se detecten aquellos estudiantes en riesgo de doble reprobación, mejor, ya que existiría mayor tiempo de intervención y reacción por parte del estudiante. Por otro lado el tiempo le otorga mayor información al modelo predictivo lo que aumenta su precisión. En un extremo, si usamos información academica de todo el primer año se logra una predicción perfecta, es decir sin errores de clasificación. En un primer momento nos propusimos estimar el modelo sólo con información del primer semestre. Los resultados no fueron satisfactorios ya que si bien nuestras predicciones fueron sustantivamente mejores que el azar, se obtuvo un alto número de errores del tipo I (falso negativos) y del tipo II (falso positivos) (estos resultados pueden ser solicitados a los autores). El mejor escenario siguiente se esquematiza en la Figura 1. Tradicionalmente, las asignaturas de primer año del Plan Común realizan tres pruebas parciales (localmente conocidas como controles) y un examen final. Tal como muestra la figura, el modelo predictivo final fue construido con información recolectada hasta la primera ronda de los controles 1 del segundo semestre. Esto deja varias semanas (un 75 % del semestre) para intervenir y dos controles y el examen final para recuperarse.

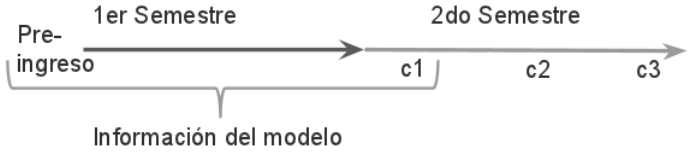

Figura 1: Tiempo de captura de información académica para el modelo

4.4. Construcción del Modelo

En cuanto al modelo predictivo, se utilizó un modelo de regresión logística en combinación con una metodología de selección de atributos, debido a la simplicidad de interpretación y utilización ampliamente aceptada. Selección de atributos puede ser considerada parte de la fase de pre-procesamiento o de minería de datos, su objetivo es encontrar el subconjunto de atributos con mayor valor predictivo, evitando así utilizar variables que agreguen ruido en la fase de entrenamiento, mejorando la predicción y acelerando así el proceso de adaptación de los modelos. Entre los enfoques más utilizados se encuentran:

- Forward Feature Selection (FS): Se comienza sin atributos en el modelo, se agregan una a una las variables y se evalúa bajo cierta métrica el desempeño de agregar cada variable, eligiéndose, de ellas, la que mejore más el desempeño (si es que hubiese mejora). El proceso se repite hasta que ninguna variable mejora el rendimiento al ser agregada.
- Backward Elimination (BE): En este enfoque se comienza con todos los atributos, luego se evalúa la eliminación de cada variable, eliminándose efectivamente la variable con mayor aumento de desempeño al ser eliminada (si es que alguna lo mejora). El proceso se repite hasta que ninguna mejora sea posible.

La metodología propuesta de selección de atributos consiste en una mezcla entre FS y BE en conjunto a una selección por frecuencia. En particular, se comienza realizando el proceso FS, tras agregar un atributo, se realiza el proceso BE. Esto con el fin de eliminar atributos ya agregados que posean mayor ruido, es decir, se pueden haber incluido nuevos atributos que, en conjunto con algunos de los atributos previamente agregados, mejoran la predicción y eliminan ruido de un atributo ya agregado.

La metodología híbrida entre FS y BE se realiza mediante validación cruzada (Cross−Validation) con cierta cantidad de conjuntos, los que van variando entre entrenamiento y validación [5]. Esto con el fin de obtener resultados representativos en cuanto al valor predictivo de cada atributo, evitando así posibles sobreajustes.

Debido a la reducida cantidad de datos que se posee en comparación a la cantidad de atributos, se combina toda la metodología previamente propuesta con una selección por frecuencia, es decir, se realiza un gran número de veces la selección de atributos y se lleva conteo de los atributos seleccionados. Finalmente se consideran como atributos seleccionados aquellos que posean una cantidad de selecciones mayor a un umbral previamente determinado.

Una vez ocurrida la selección de atributos, se utiliza un modelo de regresión logística el cual se ajusta a variables dependientes del tipo binaria (Long, 1997). La Ecuación 1 más abajo describe la función logística, donde Y representa la variable dependiente, en este caso la doble reprobación, $X_1, ..., X_n$ las variables independientes seleccionadas mediante el proceso de selección de atributos, β_0 el parámetro constante, y $\beta_1, ..., \beta_n$ los parámetros del modelo. Al estimar aquellos parámetros es posible realizar predicciones acerca de la doble reprobación basado en las variables independientes.

$$ln(\frac{Y}{1-Y}) \quad = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{1}$$

Para estimar los parámetros se utilizó la información recopilada para las cohortes de ingreso de 2010 a 2013. Luego, se usó el modelo obtenido para predecir el comportamiento del universo objetivo de la cohorte de ingreso 2014. En otras palabras, el modelo fue entrenado con las cohortes 2010-2013 y puesto a prueba con la información obtenida para la cohorte de ingreso 2014. El poder predictivo del modelo fue evaluado mediante dos reconocidos indicadores *recall* y *precision* (ver las ecuaciones más abajo), donde TP=true positive, FP = false positive, y FN=false negative , las respectivas tasas.

$$Recall \quad = \quad \frac{TP}{TP + FN} \tag{2}$$

$$Precision \quad = \quad \frac{TP}{TP + FP} \tag{3}$$

Se puede interpretar el *recall* como la tasa de los verdaderos positivos, es decir la tasa de los positivos que el modelo detecta como positivo, mientras la *precision* es la tasa de los predichos como positivo que realmente son positivo.

5. Resultados

El proceso de selcción de atributos arrojó siete variables independientes: género, tipo de establecimiento de enseñanza media, ratio de créditos reprobados, promedio controles 1 del 2do semestre menor promedio final del 1er semestre en cursos reprobados (etiqueta: C1IRS2 < FIRS1, tipo: variable binaria), diferencia con nota 4.0 del peor promedio actual en cursos reprobados en 2do semestre (C1IRS2 - 4.0, variable continua), promedio controles 1 del 2do semestre menor promedio final del 1er semestre en cursos no reprobados (C1INRS2 < C1INRS1, variable binaria), y el peor promedio controles 1 del

2do semestre menor que el peor promedio final del 1er semestre en cursos no reprobados (minC1INRS2 < C1INRS1, variable binaria). Para el caso del tipo de establecimiento de enseñanza media, partimos distinguiendo entre establecimiento público emblemático y público no emblemáticos. Luego que esta diferenciación no produjo cambios estadísticamente significativos, decidimos mantener esta variable en las tradicionales tres categorías: privado, subvencionado, y público.

La Tabla 2 muestra los resultados de la regresión logística. El test de *likelihood ratio* indica que el modelo se ajusta de buena manera a los datos (LR Chi-cuadrado = 201,62, p < 0,001). Es decir, las variables independientes tienen poder explicativo sobre el evento de doble-reprobar una asignatura. La variable que tiene el mayor poder explicativo es sin duda el ratio de los créditos inscritos reprobados. Este resultado no debiese sorprendernos. A mayor número de cursos reprobados en el primer semestre, mayor son las probabilidades de reprobar un curso por segunda vez. El poder explicativo de esta variable es tal, que se podría aplicar la heurística: si un alumno que reprueba dos o más cursos en su primer semestre, tendrá altas probabilidades de volver a reprobar al menos uno de ellosen el segundo semester. Por ejemplo, un estudiante que reprueba álgebra, cálculo, y física tiene aproximadamente cinco veces más probabilidades de doble reprobación que un estudiante que sólo reprueba una de esas asignaturas. Dos otras variables mostraron una relación estadísticamente significativa con la doble reprobación. Una de ellas es género. Un estudiante hombre tiene 88 % más probabilidades de doblereprobar que una mujer, ceteris paribus. La otra variable significativa es la diferencia entre el promedio de los primeros controles de los cursos ya reprobados y la nota de reprobación 4.0. Esto indica que aquellos estudiantes que superen la nota de aprobación en los primeros controles tienen menores probabilidades de volver a reprobar una asignatura que aquellos que no. Por ejemplo, un estudiante con promedio 3.0 en los controles 1 de las asignaturas reprobadas tiene 31 % más probabilidades de reprobar que aquel con nota 4.0, ceteris paribus.

Si bien, el resto de las variables independientes seleccionadas no son estadísticamente significativas en el modelo, el signo de los coeficientes se comporta dentro de lo esperado y es consistente con la literatura nacional. Por ejemplo, estudiantes proveniente de establecimientos de enseñanza media particular o subvencionada tienen menores probabilidades de doble reprobación que aquellos provenientes de establecimientos municipales. Los coeficientes de las tres variables binarias de rendimiento académico que no son estadísticamente significativas para el modelo también se comportan en el sentido esperado. Si el promedio de los primeros controles del segundo semestre en las asignaturas cursadas por segunda vez es menor que el promedio final de las

asignaturas reprobadas en primer semestre (C1IRS2 < FIRS1), existe una mayor inclinación a doble reprobación. Lo mismo sucede si los estudiantes bajan sus calificaciones en los controles de los ramos no reprobados en el segundo semestre con respecto al primero (C1INRS2 < C1INRS1 y minC1INRS2 < C1INRS1). Esto último es interesante ya que el modelo considera también el desempeño en aquellas asignaturas aprobadas y cursadas por primera vez.

Etiqueta	Coef.	Std. Err.	Odd Ratio
Género (hombre)	0,63**	0,30	1,88
colegio particular	-1,63	3,00	0,19
colegio subvencionado	-2,24	3,00	0,10
ratio creditos reprobados	4,41***	0,62	82,41
C1IRS2 < FIRS1	0,13	0,40	1,14
C1IRS2 - 4.0	-0,38**	0,13	0,69
C1INRS2 < C1INRS1	0,19	0,67	1,21
minC1INRS2 < C1INRS1	0,53	0,63	1,70
_cons	-3,77	3,01	
Log likelihood = -252,63			
Df = 8			
LR chi2(8) = 201,62 ***			
*p<0,1, **p<0,05, ***p<0,01			

Tabla 2: Resultado de Regresión Logística: Doble Reprobación en Primer Año (n=830)

Como se señaló en la sección anterior, el modelo fue estimado sólo con datos de las cohortes 2010-2013. Los datos de la cohorte de ingreso 2014 fueron usados para probar el poder predictivo del modelo. A cada estudiante ingresado el 2014 y con al menos una asignatura reprobada en el primer semestre, se le calculó una probabilidad de doble reprobación con información obtenida hasta la primera ronda de controles del segundo semestre. La Figura 2 muestra el resultado de esa simulación. En el eje horizontal se ubican todos los estudiantes, desde aquellos con la más alta probabilidad de doble reprobación hasta aquellos con más baja probabilidad. Dado que la variable dependiente es binaria, es necesario fijar un umbral o un porcentaje dónde aquellos con probabilidad mayor se les asigna el valor 1, la doble reprobación, y aquellos con probabilidad menor al umbral se les asigna cero o no doble reprobación. El umbral de alumnos predichos en la figura representa ese punto. El umbral fue decidido empíricamente como aquel valor de umbral donde se interceptan las curvas de sensitividad y especificidad (i.e., dónde se optimiza la correcta clasificación de casos positivos y negativos), en este caso 19 %.

Los colores representan la doble reprobación real de los estudiantes al final del segundo semestre del año 2014. Aquellas columnas en color rojo (oscuro)

representan los estudiantes que efectivamente presentaron una doble reprobación, aquellos con color verde (claro) los que no. En la figura, se evidencia que todos los que efectivamente presentaron doble reprobación, con la excepción de dos, son efectivamente predichos por el modelo. De hecho, el *recall* es 0,86, es decir el modelo clasifica correctamente a 12 de los 14 casos positivos, es decir un 86 % de las veces. Este resultado es sobresaliente en el contexto de predicciones sobre enseñanza y aprendizaje y da respaldo para generar intervenciones tempranas de apoyo. Sin embargo, al mismo tiempo el modelo clasifica incorrectamente casos negativos (i.e., falsos positivos). En total, el modelo predice 32 casos como positivos de los cuales solamente 12 son realmente positivos, lo cual da una *precision* de 37,5 %. Sin embargo, el número de falsos positivos es tolerable para el tipo de intervenciones y decisiones a tomar en base a los resultados del modelo.

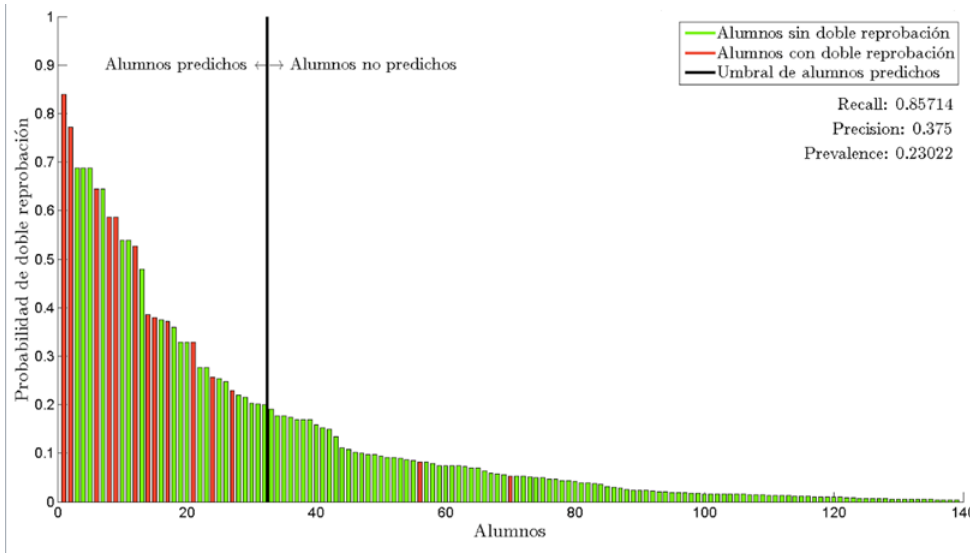

Figura 2: El Poder Predictivo del Modelo

6. Conclusiones y trabajo futuro

Este estudio tiene el objetivo de mostrar cómo herramientas de *learning analytics* pueden ser usadas para generar modelos predictivos que sirvan para apoyar a aquellos estudiantes en riesgo de deserción o de insuficientes desempeños académicos. Para tales efectos usamos datos institucionales y académicos históricos de cinco cohortes de ingreso al Plan Común de Ingeniería y Ciencias de la Universidad de Chile. Con estos datos se generó un modelo con

un alto poder prodictivo. El objetivo se centró en predecir tempranamente a aquellos estudiantes con riesgo de reprobar un mismo curso por segunda vez, lo cual los deja automáticamente en causal de eliminación. Los resultados del modelo son notables de acuerdo a su poder predictivo. Estos resultados sirvieron de base para que los directivos de la Escuela de Ingeniería y Ciencia apoyaran una serie de intervenciones de apoyo, desde comunicaciones personalizadas a los estudiantes y reforzamientos periódicos y tutorías académicas para alumnos en riesgo. Estas intervenciones se aplicarán a partir del semestre de primavera 2015.

Además el estudio abrió puertas no sólo para generar un modelo con aplicaciones prácticas, sino que también para ganar en entendimiento y generar nuevas preguntas acerca del fenómeno de la deserción y rendimiento académico. Un resultado interesante y consistente con previos estudios (p.ej., [10]), es que las estudiantes mujeres exhiben un mejor rendimiento académico que los hombres. Este resultado requiere mayor examinación, ya que otorgaría luces para la promoción de la mujer en disciplinas científicas e ingenieriles. Actualmente, tenemos en marcha un estudio que usa métodos mixtos de investigación para entender la experiencia de las estudiantes mujeres en primer año, no solo de aquellas en riesgo de caer en causal de eliminación sino que a través de todo el espectro de rendimiento académico en primer año.

Otro estudio interesante que surge a partir de este trabajo es el de aprender desde los estudiantes error tipo I, es decir aquellos a los cuales el modelo les asigna una alta probabilidad de doble reprobación pero que terminan aprobando todas sus asignaturas. ¿Qué antecedentes personales, prácticas de estudio, y actitudes determinaron el desempeño mejor de los esperado? Es una pregunta desde la cual se pueden guiar futuras políticas de intervención o la simple promoción de estrategias efectivas de estudio. En este caso, también estamos conduciendo una investigación que indaga en la experiencia de estos estudiantes para entender desde sus perspectivas cómo fue el proceso de aprendizaje durante sus primeros años de Plan Común.

En futuras investigaciones nuestro equipo está realizando esfuerzo para traer variables desde otros tipos de experiencias académicas dentro del modelo. Un área atractiva y de mayor tradición en la emergente área del *learning analytics* es aquella que estudia el uso de los sistemas en línea de gestión docente o CMS. En la FCFM el CMS local no solamente es usado para acceder información docente sino para generar discusiones, debates y otro tipo de interacciones no académicas. De algún modo entendemos que el uso de los estudiantes de estas plataforma es un indicador de su compromiso con sus planes de estudio y la vida universitaria. Estas variables nos aportarían nuevas dimensiones al análisis.

En resumen, aquí demostramos que con herramientas sencillas de *learning analytics* es posible generar modelos predictivos que permitan robustecer las decisiones curriculares y de intervención a nivel docente y administrativo. Además este estudio contribuye al entendimiento del despeño académico de los estudiantes de ingeniería y ciencias en universidades nacionales, áreas disciplinares que debiesen tener a su alcance las capacidades de usar inteligencia de datos para aprender más y ser más eficaces en los procesos educativos.

Agradecimientos: Este trabajo fue financiado por el proyecto Basal "Diseño de un sistema de información para el monitoreo, evaluación y mejoramiento continuo de la docencia de pregrado"(UCH1298) y el Instituto Sistemas Complejos de Ingeniería (ICM: P-05-004-F, CONICYT: FB016).

Referencias

[1] A. Acharya y D. Sinha. Application of feature selection methods in educational data mining. *Journal of Computer Applications*, 103(2):34–38, 2014.

[2] C. Acuña Veliz. Acceso y deserción en la educación superior: caso aplicado a Chile. *Tesis de Magíster. Universidad de Chile, Santiago, Chile*, 2012.

[3] F. Angulo y E. Sergio. Modelo para la automatización del proceso de determinación de riesgo de deserción en alumnos universitarios. *Tesis de Magíster. Universidad de Chile, Santiago, Chile*, 2012.

[4] P. Baepler y C. J. Murdoch. Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning*, pages 1–9, 2010.

[5] B. Baesens. Analytics in a big data world. *John Wiley and Sons*, 2014.

[6] R. S. Baker. Educational data mining: An advance for intelligent systems in education. *IEEE Intelligent Systems*, pages 78–82, 2014.

[7] J. P. Bean. Student attrition, intentions, and confidence: Interaction effects in a path model. *Research in Higher Education*, 14(4):291–320, 1982.

[8] M. Berland, R. S. Baker, y P. Blikstein. Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, 19(1-2):205—220, 2014.

[9] A. F. Cabrera, A. Nora, y M. B. Castañeda. College persistence: Structural equations modeling test of an integrated model of student retention. *Journal of Higher Education*, 64(2):123–139, 1993.

[10] S. Celis. Student attrition and student time-to-degree at a selective engineering school in Chile. *Documento interno, Escuela de Ingeniería y Ciencias, Universidad de Chile*, 2012.

[11] R. Chen. Institutional characteristics and college student dropout risks: A Multilevel Event History Analysis. *Research in Higher Education*, 53(5):487–505, 2012.

[12] R. Chen y S. L. DesJardins. Exploring the effects of financial aid on the gap in student dropout risks by income level. *Research in Higher Education*, 49(1):1–18, 2007.

[13] C.J. Díaz. Factores de deserción estudiantil en ingeniería: Una aplicación de modelos de duración. *Información Tecnológica*, 20(5):129–145, 2000.

[14] Centro de Microdatos. Estudio sobre causas de la deserción universitaria. *Departamento de Economía, Universidad de Chile*, 2008.

[15] S. L. DesJardins, D. A. Ahlburg, y B. P. McCall. An event history model of student departure. *Economics of Education Review*, 18(1):375–390, 1999.

[16] E. Himmel. Modelos de análisis de la deserción estudiantil en la educación superior. *Calidad de la Educación*, 17:91–107, 2002.

[17] P. Jia y T. Maloney. Using predictive modelling to identify students at risk of poor university outcomes. *Higher Education*, 70:127–149, 2014.

[18] I. Johnson. Enrollment, persistence and graduation of in-state students at a public research university: Does high school matter? *Research in Higher Education*, 49(8):76–793, 2008.

[19] T. Larroucau. Estudio de los factores determinantes de la deserción en el sistema universitario chileno. *Tesis de Magíster. Universidad de Chile, Santiago, Chile*, 2013.

[20] S. A. Lesik. Do developmental mathematics programs have a causal impact on student retention? an application of discrete-time survival and regression discontinuity analysis. *Research in Higher Education*, 48(5):583–608, 2007.

[21] J. Luan, T. Kumar, S. Sujitparapitaya, y T. Bohannon. Exploring and Mining Data. in: R.D. Howard, G.W. McLaughlin, W.E. Knight (eds.). *The Handbook of Institutional Research. San Francisco, CA: Jossey-Bass*, pages 478–501, 2012.

[22] T. Martin y B. Sherin. Learning analytics and computational techniques for detecting and evaluating patterns in learning: An introduction to the special issue. *Journal of the Learning Sciences*, 22(4):511–520, 2013.

[23] SIES Ministerio de Educación. Retención de primer año en educación superior. programas de pregrado, 2014.

[24] SIES Ministerio de Educación. Retención en educación superior con perspectiva de género, 2014.

[25] A. Mizala, T. Hernández, y M. Makovec. Determinantes de la elección y deserción en la carrera de pedagogía. *Proyecto FONIDE N° F511059*, 2011.

[26] E. T. Pascarella y P. T. Terenzini. How college affects students. *San Francisco, CA: Jossey-Bass*, 2, 2005.

[27] R. Pea. The learning analytics workgroup: A report on building the field of learning analytics for personalized learning at scale. 2014.

[28] R. Rolando, J. Salamanca, A. Lara, y C. Blanco. Deserción y reingreso a la educación superior en Chile: Análisis de la cohorte 2008. *SIES, Ministerio de Educación*, 2012.

[29] I. Roll y P. H. Winne. Understanding, evaluating and supporting self-regulated learning using learning analytics. *Journal of Leaning Analytics*, 2(1):7–12, 2015.

[30] C. Romero y S. Ventura. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[31] V. Santelices, X. Catalán, C. Horn, y D. Kruger. Determinantes de deserción en la educación superior chilena, con Énfasis en efecto de becas y créditos. *Proyecto FONIDE N° F611103*, 2013.

[32] G. Siemens y R. S. J. Baker. Learning analytics and educational data mining : Towards communication and collaboration. 2012.

[33] L. D. Singell y G. R. Waddell. Modeling retention at a large public university: Can at-risk students be identified early enough to treat? *Research in Higher Education*, 51(6):546–572, 2010.

[34] V. Tinto. Stages of student departure: Reflections on the longitudinal character of student leaving. *Journal of Higher Education*, 59(4):438–455, 1988.

[35] C. Ye y G. Biswas. Early prediction of student dropout and performance in moocs using higher granularity temporal information. *Journal of Learning Analytics*, 1(3):169–172, 2014.