UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

EXPLORANDO LA RELACIÓN ENTRE LA ACTIVIDAD EN LÍNEA Y EL RENDIMIENTO ACADÉMICO DE ESTUDIANTES EN PRIMER AÑO DE INGENIERÍA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL ELÉCTRICO

FRANCISCO NICOLAS CATRILEO HERRERA

PROFESOR GUÍA:
SERGIO CELIS GUZMÁN

MIEMBROS DE LA COMISIÓN:
MARCOS ORCHARD CONCHA
ANDRÉS CABA RUTTE

SANTIAGO DE CHILE
2017

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO
POR: FRANCISCO NICOLAS CATRILEO HERRERA
FECHA: 17/05/2017
PROF. GUÍA: SERGIO CELIS GUZMÁN

EXPLORANDO LA RELACIÓN ENTRE LA ACTIVIDAD EN LÍNEA Y EL RENDIMIENTO ACADÉMICO DE ESTUDIANTES EN PRIMER AÑO DE INGENIERÍA

Actualmente, la cantidad de datos que se generan mundialmente crece de manera exponencial. Además existen muchas actividades que se han beneficiado de la capacidad de generar y tratar datos para crear valor socioeconómico y ambiental. En este contexto surge el *Learning Analytics (LA)* que incluye la generación y el tratamiento de datos para el ámbito educativo. Por una parte, las instituciones de educación superior cuentan con mucha información de distinta naturaleza, por ejemplo, información de admisión, datos demográficos y académicos o registros de los *Learning Management Systems (LMS)*. Por otra parte, las instituciones procesan esta información para la toma de decisiones. En este sentido, la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile busca mejorar su actividad docente, a partir de datos en sus sistemas y el contraste de esta información con la Teoría del Aprendizaje.

Dentro de las teorías más reconocidas sobre el aprendizaje en educación superior está la desarrollada por Marton y Säljö[22], y Asikainen et al.[4] la cual propone tres aproximaciones del estudiante al aprendizaje: profundo, superficial y estratégico. En general, las instituciones aspiran a que el estudiante tenga un aprendizaje profundo. Entonces la presente memoria busca establecer un puente entre la información disponible de las instituciones y la Teoría del Aprendizaje. En este sentido, el trabajo realizado tiene como objetivo saber si la información provista por los *LMS* sumado a la información académica y de admisión permite tener una buena estimación del rendimiento académico y saber si se refleja lo propuesto en la Teoría del Aprendizaje. Entonces se implementa un modelo matemático de dos formas: una solamente con antecedentes y notas de los estudiantes, y otra con la información anterior más los datos provistos por el *LMS* de la Facultad (llamado U-cursos). El modelo implementado consiste en un preprocesamiento, en la implementación de *Similarity Based Modeling (SBM)* y un post-procesamiento. En primer lugar, el preprocesamiento consiste en establecer métricas que indiquen la naturaleza temporal y cuantitativa de los *logs* de U-cursos, acotar las variables a valores $\in [0, 1]$ y seleccionar las variables que se utilizarán mediante *Partial Least Squares* y *Principal Components Analysis*. Luego, se implementa un predictor de nota final utilizando *SBM*. Por último, se clasifican las predicciones entre aprobados y reprobados.

En efecto, en el proceso de la selección de variables se descartan aquellas como región de procedencia, tipo de establecimiento y PSU de lenguaje. También variables relacionadas con la actividad *online* correspondiente al módulo material alumnos. Asimismo, dentro de las variables con mayor incidencia destacan desempeños académicos anteriores, género, año de egreso de enseñanza media y actividad en el foro de U-cursos. Por otro lado, el clasificador mejora considerablemente su desempeño (un 15 % en términos de *recall*) cuando se incluyen los datos de U-cursos, lo que implica que éstos permiten tener mayor información útil respecto a lo que el estudiante hace fuera del aula. En otras palabras, este modelo sienta las bases para caracterizar a estudiantes según el tipo de aproximación al aprendizaje.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

La cantidad de datos que se están generando es cada vez mayor y crece de manera exponencial. Tanto es así, que en IBM se dice que el 90 % de los datos del mundo han sido generados en los últimos dos años[18]. A este enorme conjunto de datos de fuentes diferentes, de naturaleza y confiabilidad diversa, se le conoce como *big data*. Además, el recolectar esta enorme cantidad de datos sumado a convertirlos en información pertinente para el uso y la toma de decisiones trae oportunidades. Dentro de las oportunidades que trae el *big data* están la predicción de fallas, lo cual implica un mantenimiento preventivo óptimo o la previsión de insumos necesarios, optimizando los gastos de las instituciones. De hecho, existe evidencia que muestra que los datos que están de forma abierta[1], sumado a nuevos métodos para la colección y tratamiento de datos puede ahorrar dinero y crear valor económico, social y medio-ambiental[24]. Estos ahorros y nuevos valores han sido percibidos por variadas industrias, áreas del conocimiento, gobiernos e instituciones. Bajo este contexto se presenta una oportunidad para las instituciones de educación superior, la cual consiste en otorgar valor a sus datos con el objetivo de mejorar los procesos de aprendizaje de sus estudiantes. En este sentido, existe un campo del conocimiento en el cual el uso de los datos orientado a la educación juega un rol importante. Este campo es el *Learning Analytics (LA)*, el cual tiene el objetivo de analizar y mejorar las experiencias de aprendizaje mediante la recolección e interpretación de datos seguidos de una acción[9]. En particular, el uso de técnicas de *machine learning* para propositos educacionales *(EDM)* es un campo prometedor. Pues, las instituciones cuentan generalmente con datos como cursos virtuales, e-learning log files, registros de *Learning Management Systems (LMS)*, datos demográficos y académicos e información de la admisión, los cuales pueden ser procesados con técnicas de *machine learning* con el fin de obtener interpretaciones, avisos, sugerencias, entre otros[19]. Por otro lado, se ha observado una tendencia mundial hacia el desarrollo de este campo, ya que ha permitido optimizar los procesos de aprendizaje y enfrentar problemas como el abandono de carreras universitarias[3].

Las aspectos positivos de utilizar *LA* en las instituciones pueden ser percibidas por los actores más importantes en el proceso educativo: los profesores, los estudiantes y administra-

[1]Se refiere al término *Open Data*, que propone la idea de que algunos datos estén disponibles para el uso de todas las personas.

dores de la institución. En primer lugar, los profesores pueden tener a disposición una imagen más completa de lo que está ocurriendo en el curso, ya que actualmente pueden observar solo lo que los estudiantes hacen en las clases, aulas y evaluaciones formales. En consecuencia, los profesores pueden tener un indicio de cómo los estudiantes van llevando su curso, del compromiso entre alumnos y curso, de cuáles estudiantes están en riesgo de abandonar o con falta de motivación y/o de cuales estudiantes tienen un interés académico importante. En este sentido, el profesor podría por ejemplo, hacer intervenciones para comprender fenómenos como la desmotivación. Además, el profesor podría tener una herramienta adicional para poder evaluar innovaciones o cambios que haya decidido aplicar[14]. En segundo lugar, los estudiantes podrían observar sus predicciones para saber si deben cambiar de estrategia en el estudio o buscar ayuda. En tercer lugar, las autoridades y administrativos de la institución que utilice *LA* puede implementar políticas que mejoren su actividad. Por ejemplo, la *Ball State University* implementó un sistema para identificar estudiantes en riesgo de deserción que ha permitido mejorar la tasa retención en un 4 % en 3 años. Otro ejemplo es el caso de la Universidad de Alabama, en donde se impuso que todos los estudiantes de primer año deben vivir en el campus, ya que se descubrió que quienes viajaban todos los días tenían un alto riesgo de desertar[8].

Un caso especial de las fuentes de información para el desarrollo del *LA* corresponde a los registros generados por los *LMS*. Antes de todo, los *LMS* son una herramienta para administrar recursos de apoyo al aprendizaje, en particular en cursos dictados en universidades. Además, tienen varias funcionalidades, tales como, poner a disposición material de estudio, planificar actividades en un calendario, administrar tareas y notas, discusión en foros, enviar correos, entre otros. Según el ranking de las mejores 200 herramientas para el aprendizaje, elaborado por el *Centre for Learning and Performace Technologies*, los *LMS* más importantes son: Moodle, Canvas, Google Classroom, Edmodo, Blackboard, Desire2Learn y Sakai[16]. Entonces, estos programas elaboran registros o *logs* respecto de toda la actividad que se realiza dentro del *software*, entregando información valiosa para el *LA*.

En este caso, la institución que busca desarrollar el *LA* en su actividad es la Universidad de Chile, específicamente la Facultad de Ciencias Físicas y Matemáticas (FCFM). Esta Facultad tiene una población estudiantil de 5500 en donde el 23 % corresponde a mujeres y un tercio corresponde a estudiantes que no provienen de la región metropolitana[2][12]. Estos estudiantes son seleccionados del 3 % superior de la enseñanza media de acuerdo a la Prueba Nacional de Selección Universitaria(PSU). También, la FCFM la componen cerca de 1200 estudiantes de postgrado y 220 profesores de jornada completa, de los cuales un 97 % posee un grado de doctor[7]. Por otro lado, en esta Facultad se albergan 9 programas de ingeniería civil de pregrado, 3 licenciaturas en ciencias y el programa de geología. Para seguir cualquiera de los programas anteriores, los estudiantes deben cursar 4 semestres de Plan Común, en el cual se asientan conocimientos básicos de matemáticas y ciencias naturales. Estos semestres tienen una carga académica de 30 SCT (Sistema de Créditos Transferibles), lo que equivale a 50 horas de trabajo semanal durante 15 semanas. Asimismo, durante los dos primeros años los estudiantes asisten a cátedras que en su mayoría tienen del orden de 100 estudiantes. La mayoría de las cursos consisten en una o dos cátedras a la semana dictadas por el profesor encargado del curso y de una clase auxiliar en donde se resuelven problemas similares a los que

[2]Región en donde se encuentra la Facultad. Además corresponde a la región con mayor población y actividad económica del país.

se evaluarán. En general, las evaluaciones tienen 3 instrumentos: controles, tareas o ejercicios y examen. Primero, se tienen 3 controles distribuidos a lo largo del semestre, exceptuando los cursos de matemáticas de primer semestre que tienen 6 controles (1 cada dos semanas). Además, durante el primer semestre los estudiantes de primer año tienen la opción de dar un control recuperativo que permite reemplazar el peor control hasta la fecha. Segundo, las tareas o ejercicios van variando según el profesor a cargo y no hay regla general para estas evaluaciones. Tercero, el examen se realiza a final de año y tiene un 40 % de peso dentro del resto de las evaluaciones. Existe la posibilidad de no hacer este examen para los estudiantes que han tenido un rendimiento alto en sus controles. En general, se reconoce que la FCFM tiene una alta exigencia académica. De hecho los estudiantes aprueban en general el 85 % de sus cursos.

Por otro lado en la FCFM se utiliza un *LMS* llamado U-cursos, el cual proporciona los datos para esta Memoria de Título. Esta plataforma ha sido desarrollado desde la década de los 90 al interior de la Facultad, con el objetivo de organizar los recursos educativos utilizados en sus cursos. Luego, el sistema ha tenido un gran crecimiento, extendiéndose a otros organismos de la Universidad y también a otras universidades. En cuanto a la propiedad de los datos que se generan en esta solución TI, la Universidad es proprietaria.

Además, la Facultad ha tenido otros esfuerzos en *LA*. Un caso es el modelo propuesto por Celis et al.[7], en donde se propone un modelo analítico para identificar tempranamente a estudiantes de primer año en ingeniería que caerán en causal de eliminación. En este estudio se utilizan datos socio-demográficos, académicos y datos de admisión. El presente trabajo integra esas fuentes de información y los registros de U-cursos.

Cabe considerar que la transición desde la educación media hacia la educación superior tiene demandas significativas para los estudiantes, ya que la vida universitaria es más exigente y estresante. Además, requiere mayores niveles de independencia, autonomía, iniciativa y regulación personal *(self-regulation)*. De hecho, se estima que cerca de la mitad de los estudiante que desertan en sus respectivos programas lo hacen en primer año[1]. Entonces, es importante que los estudiantes tengan un buen rendimiento en primer año, ya que éste se relaciona con la autoeficacia[3] y el optimismo. Además, la autoeficacia y el optimismo se relacionan con la persistencia, la tenacidad y los logros en el ámbito académico[10]. Sumado a lo que dice Martin et al.[10], Tinto[28] menciona que los primeros seis meses en la universidad es un período altamente importante en la persistencia del estudiante. También señala que completar el primer año es más de la mitad de la «batalla» de la persistencia en los estudios de pregrado. Es por esta razón que la presente memoria se enfoca en el estudio del primer semestre.

Por otro lado, se ha avanzado en teorías de aprendizaje, que permiten entender diferencias en rendimiento y éxito académico. Marton y Säljö[22], y Asikainen et al.[4] han desarrollado las teorías más aceptadas por la comunidad académica. En estos trabajos, se propone que los estudiantes universitarios tienen formas diferentes de llevar su proceso de aprendizaje: Aprendizaje Profundo, Aprendizaje Superficial y Aprendizaje Estratégico. En primer lugar,

[3]Se entiende por autoeficacia como la creencia del sujeto en la capacidad de si mismo para organizar y ejecutar acciones requeridas para adquirir un logro[30].

el Aprendizaje Profundo se caracteriza por un entendimiento amplio, en donde el estudiante tiene un interés académico por la materia. En segundo lugar, el Aprendizaje Superficial se describe como un acercamiento trivial hacia los tópicos, mediante la memorización de palabras (símbolos). En tercer lugar, el Aprendizaje Estratégico se relaciona con la forma en que el estudiante organiza su estudio dependiendo del tipo de evaluación del curso. Sin embargo, los estudios mencionados anteriormente no toman en consideración los datos disponibles que tiene la institución. En otras palabras, para cada estudio se han debido hacer entrevistas o encuentas, generando nueva información. Una manera de complementar, verificar o entender mejor las teorías es con la información contenida en los *LMS*, que entrega indicios sobre el comportamiento del estudiante fuera del aula. En consecuencia, la presente memoria busca establecer un puente entre la gran cantidad de datos que dispone la facultad y estas teorías del aprendizaje.

En resumen, existe una tendencia mundial hacia mejorar la generación y procesamiento de datos, ya que esto ha beneficiado variados sectores de la industria y servicios públicos. En este contexto, las instituciones de educación promueven el desarrollo del *Learning Analytics*, porque el *LA* puede otorgar herramientas para mejorar la actividad de las universidades. En particular, la FCFM ha comenzado a potenciar este campo. También, se hace hincapié en el primer año, ya que la transición de la educación media a la educación superior está relacionada con la persistencia, la motivación, autoeficacia y las perspectivas académicas en la universidad. Además, la memoria busca establecer una relación entre las teorías del aprendizaje y el *LA* a través de un modelamiento empírico que permita la optimización de recursos institucionales.

1.1. Objetivos

Objetivo General

Crear un modelo empírico, a partir de datos, que permita entender de mejor forma los procesos de aprendizaje de estudiantes de primer año de ingeniería, a partir de datos socio-demográficos, académicos y registros de *LMS* que posee la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile.

Objetivos Específicos

1. Inferir una relación entre el análisis de datos y las teorías en educación.
2. Relacionar registros de un *LMS* con datos de rendimiento académico y otros antecedentes individuales de los estudiantes.
3. Implementar una técnica de análisis de datos que no haya sido utilizada en *Learning Analytics* de la FCFM.
4. Proponer estrategias para la interpretación y el uso de *LA* en la FCFM.

1.2. Alcances

La presente Memoria de Título está focalizada en el diseño e implementación de un modelo para predecir el rendimiento académido en estudiantes de primer año de la FCFM. Para lograr esto, se utilizará una herramienta matématica del campo del reconocimiento de patrones llamada *SBM (Similarity Based Modeling)*. Además se utilizarán las técnicas *PLS (Partial Least Squares)* y *PCA (Principal Component Analysis)* para la selección y preprocesamiento de variables.

Por otro lado, la información utilizada para construir estos modelos consiste en datos de estudiantes que ingresaron en 2013 y 2014. Dentro de los datos de cada estudiante se tienen las notas, los antecedentes y la actividad online de la plataforma *LMS (Learning Management System)* de U-cursos. Además, se toma en cuenta solamente el curso Introducción al Cálculo - MA1001, porque es considerado muy importante ya que en caso de que un estudiante reprobare, éste estará imposibilitado de tomar tres cursos de la malla curricular el próximo semestre. Además, Introducción al Cálculo es considerado como un curso difícil, de hecho este curso tiene una reprobación del 25 % el año 2013. También, cursos similares se encuentran en casi todos los programas de ingeniería mundialmente.

En primer lugar, las notas corresponden a las actas enviadas en cada uno de los cursos. En donde, la mayoría de las actas tiene la información de los controles a lo largo del semestre, junto con el examen y notas de ejercicios o tareas. En segundo lugar, los antecedentes corresponden a la información que tienen los estudiantes al ingresar a la facultad, por ejemplo: Región de procedencia, tipo de establecimiento, puntaje PSU, entre otros. Por último, la actividad online de U-cursos corresponde a todo lo que se ha realizado dentro de la plataforma que concierne al curso, tal como lectura, respuestas y publicaciones de mensajes en foro, descarga de material docente o descarga de material alumnos.

Entonces, el trabajo aquí presentado consiste en implementar este modelo *SBM*. Para lograr esto, se debe recopilar esta información de diferentes fuentes y programar.

1.3. Estructura General

La presente Memoria de Título cuenta con 7 capítulos. El primer capítulo introduce el tema, plantea objetivos y se definen los alcances. Luego, en el Capítulo 2 se presenta la revisión bibliográfica considerando conceptos de LA, teorías del aprendizaje y herramientas matemáticas. Después, en el Capítulo 3 se muestra la metodología con que se aborda el tema. En el capítulo siguiente, se expone la implementación del problema. Posteriormente en el Capítulo 5 se exhiben los resultados obtenidos. Luego, en el Capítulo 6 se plantea una discusión a partir de los resultados obtenidos. Por último, se presentan las conclusiones de la memoria.

Capítulo 2

Revisión Bibliográfica

2.1. Learning Analytics

Actualmente el interés por *LA* es creciente dentro de las instituciones de educación superior. De hecho, las instituciones usan LA con el objetivo de estudiar y mejorar la retención de estudiantes, el éxito académico y la reducción del tiempo para la graduación de estudiantes. Entonces, existe un gran oportunidad para el desarrollo de esta área[3]. Sin embargo, la inversión en *LA* no corresponde a una parte importante de la inversión total en las instituciones. *LA* no es un campo de investigación específico, ni uno nuevo, sino mas bien es la unión de distintos conceptos relacionados entre sí. De hecho, el LA incluye el *Academic Analytics (AA), Educational Data Mining, Recommender Systems* y *Personalized Adaptive Learning.* Primero, el *Academic Analytics* es la aplicación del *Business Intelligence* en el campo de la educación superior, es decir, utilizar datos de una (o varias) institución con técnicas estadísticas para asistir a la misma en la toma de decisiones. Segundo, *Educational Data Mining (EDM)* se refiere a la aplicación de técnicas de minería de datos en el campo de la educación. Alguno ejemplos de estas herramientas son: *clustering*, clasificación o *association rule mining.* Tercero, *Recommender Systems* se refiere al uso de información del usuario para recomendar ítems en los cuales el usuario podría estar interesado. Cuarto, *Personalized Adaptive Learning* se refiere a sistemas automáticos en donde se guía al estudiante a través del material de aprendizaje y/o se recomienda al mismo a través de los cursos. En conclusión, se considera que LA corresponde a todas las áreas de investigación y desarrollo de métodos para analizar y detectar patrones con información recogida en los ambientes educacionales para luego tomar acciones que enriquezcan la experiencia de aprendizaje[9]. En este sentido, *EDM* y *AA* tienen una importancia especial en el marco de esta memoria.

Dentro de algunos ejemplos en los que se ha utilizado *LA* se encuentra el trabajo de Kotsiantis[19], en donde se citan varios casos de usos de técnicas de *machine learning* y se presenta un estudio de caso en el cual se predice las notas de los estudiantes. En los recursos citados por Kotsiantis se incluyen métodos de clasificación, regresión, reglas de asociación, *clustering*, minería de secuencias y minería *WEB*. En primer lugar, para la clasificación se utiliza *decision tree*, *link analysis* y *decision forest* para analizar preferencias en cursos, tasas de aprobación de cursos y optimizar las secuencias de aprendizaje. Igualmente, se utiliza *abduc-*

tive network modeling para identificar los ítems en un control que aportan mayor información sin perder mucha exactitud. En otras palabras, reducir las preguntas en una evaluación, sin alterar los puntajes de los estudiantes (lo menos posible). Asimismo, se utiliyan *feed-forward neural networks*, *support vector machines* y *ARTMAP*. En segundo lugar, los métodos de regresión se han utilizado para realizar predicciones sobre las notas finales de los estudiantes. En estas predicciones se observa que los puntajes en evaluaciones son medianamente predictivos, pero cuando esta información se combina con la información de los *logins* del *LMS* se logra una mayor capacidad predictiva[6][21]. También, se identifica que variables como el total de correos electrónicos enviados, total de mensajes posteados, total de evaluaciones completadas pueden explicar mas del 30 % de la variación de la nota final del estudiante. En tercer lugar, el uso de reglas de asociación se utiliza para identificar confusiones recurrentes en los estudiantes. Asimismo se ha utilizado para identificar estudiantes que necesitan ayuda adicional. También, se ha combinado con la teoría de conjuntos difusos para construir reglas de asociación difusas. En cuarto lugar, los algoritmos de *clustering* se han utilizado para analizar colaboración dentro de los *learning environments*. Además, se ha utilizado para evaluar e identificar factores en el desempeño de los estudiantes en un *learning environment* basado en *WEB*. Igualmente se ha utilizado para implementar sistemas de recomendación en cursos de inglés. Dentro de las técnicas de *clustering* mencionadas se tiene *SOM* y *K-means*. En quinto lugar, *sequential pattern mining* se ha utilizado para recomendar secuencias de uso de recursos para los usuarios. Además se ha utilizado para identificar el comportamiento de grupos exitosos. Por último, la minería *WEB* se ha utilizado para agrupar textos con un tópico similar y realizar recomendaciones de recursos.Otro ejemplo es el de Sael, Marzak y Behja[27] quienes aplican técnicas de *clustering* y *association rule mining* para analizar el comportamiento de estudiantes contribuyendo con las evaluaciones de aprendizaje y para mejorar la estuctura de un contenido *SCORM*[1]. Esto implica que el *LA* puede ser utilizado como una herramienta adicional de evaluación de cursos e innovaciones dentro de estos.

En conclusión, *LA* es una forma de estudiar el aprendizaje en la cual se utilizan recursos de las Tecnologías de la Información. Además, se presenta como una oportunidad para enriquecer la actividad de los actores de las instituciones de educación superior. Por ejemplo, al obtener predicciones de la nota final de estudiantes se pueden realizar intervenciones tempranas o al analizar las preferencias hacia cursos y tasas de aprobación se pueden optimizar los recursos institucionales en términos de la cantidad de cursos que se dictan. Sin embargo, la implementación de *LA* requiere la recolección y ordenamiento de datos, además de la formación de equipos que se dediquen a la investigación y desarrollo de aplicaciones. A continuación, se presenta una aproximación teórica al estudio del aprendizaje.

[1] *SCORM* es un conjunto de estándares técnicos para *e-learning software*, que tiene como objetivo la integración fácil de diferentes programas.

2.2. Teorías del Aprendizaje

Aprendizaje Profundo, Aprendizaje Superficial y Aprendizaje Estratégico

Marton y Säljö [22] proponen que existen dos tipos de aproximarse al aprendizaje: Aprendizaje Profundo y Aprendizaje Superficial. Para explicar esto, se elabora un experimento en el que estudiantes leen un pasaje, para luego responder preguntas sobre el significado del texto y de su forma de abordar la lectura. Primero, el Aprendizaje Profundo se caracteriza por una comprehensión vasta del tema, con una interpretación cercana a lo que ha querido transmitir el autor del texto leído. En esta forma de aprendizaje el estudiante dirige su atención hacia la intención del autor. Además, coincide con que estos estudiantes tienen un interés académico, entendiéndose como la atracción del estudiante hacia el contenido. Segundo, el Aprendizaje Superficial se compone simplemente de memorización para hechos o información sin una comprensión de la intención del autor o de conceptos abstractos. En este caso el estudiante dirige su atención hacia el texto en sí (los carácteres), tal como lo haría para memorizar un número telefónico, sin darle sentido a las palabras que lee. Además, estos estudiantes suelen no tener un gran interés académico.

Por otro lado, Asikainen, Parpala, Virtanen y Lindblom-Ylänne[4] complementa a Marton y Säljo proponiendo un tercer acercamiento al aprendizaje: el Aprendizaje Estratégico. En esta forma de abordar el aprendizaje se enfatiza la motivación en el estudio, en donde la intención del estudiante es obtener la mejor calificación estudiando según el tipo de evaluación. Por otro lado, la motivación está fuertemente ligada al acercamiento de aprendizaje del estudiante. De hecho, la motivación intrínseca del estudiante conduce a un afinidad por el aprendizaje profundo[15]. También varios estudios han mostrado que la auto-regulación y el estudio organizado ha mostrado tener relación con una aproximación profunda del aprendizaje. Entendiendo la auto-regulación como la habilidad de ponerse metas en el aprendizaje.

Además, varios estudios han mostrado una relación positiva entre Aprendizaje Profundo y el éxito en el estudio[25], junto con una relación negativa entre *surface learning* y éxito en el estudio. Por otro lado, en [26] se encuentra una relacion entre estudio organizado y éxito en el estudio pero sin relación con el acercamiento profundo.

Como resultado en la investigación de Asikainen et al.[4], se obtiene que solamente los estudiantes que no tienen un buen manejo del tiempo pueden tener un logro pobre, de hecho todos los estudiantes con una aproximación profunda pero con un manejo del tiempo mediocre tienen un logro bajo[2].

En conclusión, se supone que un estudiante con un enfoque profundo tiene un comportamiento constante a lo largo del semestre. En otras palabras, consulta el material de estudio desde el inicio del semestre y lo repasa constantemente, además, resuelve sus dudas inmediatamente, discute y comparte material sobre lo tratado en el curso. Luego, se asume que los estudiantes con un enfoque estratégico tienden a organizar su tiempo de estudio en función de sus evaluaciones. Entonces, se esperaría que estos alumnos consulten el material de estudio y resuelvan sus dudas con algún grado de anticipación, pero no constantemente. Por

último, se considera que los estudiantes con un aprendizaje superficial consulten el material de estudio y resuelvan sus dudas solamente durante un período corto e inmediatamente antes de la evaluación.

2.3. Herramientas de Análisis de Datos

2.3.1. Sistemas de reconocimiento de patrones

En los sistemas de reconocimiento de patrones se tiene un ciclo similar para todas las aplicaciones[2]. Este ciclo consiste en: recopilación de datos, extracción y selección de características, elección del modelo, entrenamiento y validación[13]. En primer lugar, se juntan y preparan las bases de datos para que sean tratadas, p. ej. registros del *LMS* o mediciones de un sensor en el tiempo. En segundo lugar, se extraen las características o propiedades que se estimen interesantes a partir de las bases de datos para cada observación, p. ej. número de *logins* en la plataforma *LMS* por estudiante. En este caso, cada estudiante corresponde a una «observación». Entonces, se crea un conjunto de observaciones con sus respectivas características. Además, en esta etapa se puede reducir el número de propiedades (selección de características). En tercer lugar, se elige un modelo matemático para tratar el problema. En cuarto lugar, se realiza el entrenamiento. Para realizar el entrenamiento se debe dividir el conjunto creado en la extracción y selección de características en dos conjuntos de manera aleatoria: Conjunto de entrenamiento y conjunto de validación. En general, el conjunto de entrenamiento es más grande que el conjunto de validación. Entonces, el modelo matemático seleccionado se ajusta a partir de los datos del conjunto de entrenamiento. En quinto lugar, la evaluación consiste en operar el modelo «entrenado» con el conjunto de validación y medir el rendimiento comparando la clasificación hecha por el modelo con la clasificación real, por ejemplo, mediante la matriz de confusión, el *recall* y la precisión.

2.3.2. Selección de variables mediante *Partial Least Squares*

Antes de entender los distintos métodos de selección de variables en *Partial Least Squares (PLS)* vale la pena comprender el algoritmo *PLS regression (PLSR)*. Entonces, *PLS* permite la modelación de sistemas proponiendo una reducción de dimensionalidad del problema, así como la maximización de la covarianza entre la matriz de variables explicativas $X_{(n,m)}$ y la matriz de respuestas $Y_{(n,p)}$. En este caso m es el número de variables, n es el número de observaciones y p es el número de variables de respuesta. Se supone una relación lineal tal que $Y = \alpha + X\beta + \varepsilon$, en donde α y β son los parámetros de regressión y ε es un término de error[23][20].

Inicialmente las variables se encuentran centradas y escaladas. Además se supone un paramétro A tal que $A \leq m$ que corresponde al numero de componentes relevantes para la

[2]En la presente Memoria de Título se considera solamente el reconocimiento de patrones a través del aprendizaje supervisado.

predicción. Entonces para un parámetro $a = 1, 2, ..., A$ el algoritmo *PLSR* sigue[23]:

1. Se computan los pesos :
$$w_a = X'_{a-1} Y_{a-1} \tag{2.1}$$
Los pesos definen la dirección en el espacio abarcada por X_{a-1} de máxima covarianza con Y_{a-1}. Luego se normalizan estos pesos.
2. Después se computa el vector *score* con:
$$t_a = X_{a-1} w_a \tag{2.2}$$
3. Posteriormente se computan las $X-$cargas p_a con:
$$p_a = X'_{a-1} \frac{t_a}{t'_a t_a} \tag{2.3}$$
4. De manera similar se calculan las $Y-$cargas q_a:
$$q_a = Y'_{a-1} \frac{t_a}{t'_a t_a} \tag{2.4}$$
5. Seguidamente se disminuye X_{a-1} e Y_{a-1} al sustraer la contribución de t_a:
$$X_a = X_{a-1} - t_a p'_a \tag{2.5}$$
$$X_a = Y_{a-1} - t_a q'_a \tag{2.6}$$
6. Por último, si $a < A$ volver a 1.

Se debe almacenar los pesos, los *scores* y cargas en las matrices/vectores $W = [w_1, w_2, ..., w_A]$, $T = [t_1, t_2, ..., t_A]$, $P = [p_1, p_2, ..., p_A]$ y $Q = [q_1, q_2, ..., q_A]$. Entonces los *PLSR*-estimadores son $\hat{\beta} = W(P'W)^{-1}Q$ y $\hat{\alpha} = \bar{y} - \bar{x}\beta$.

Según como se define la selección de variables en *PLSR* se pueden categorizar en 3 tipos diferentes: métodos de filtro, métodos de envoltorio y métodos embeídos. En primer lugar, los métodos filtro utilizan las salidas del algoritmo *PLRS* para solamente identificar un subconjunto de variables significativas. Además, constituye el método más simple. En segundo lugar, los métodos de envoltorio vuelven a integrar las variables identificadas al utilizar un método de filtro. En tercer lugar, los métodos embeídos integran la identificación de variables como parte del algoritmo *PLSR*.

Para lo que sigue de la memoria solamente se mencionará un método de filtro, debido a su simpleza y a que es el método seleccionado. Primeramente, se ajusta el modelo *PLRS* a los datos, luego la selección de variables se hace mediante la imposición de un umbral a una medida relevante del modelo. Entonces, la medida utilizada en la presente memoria son los pesos w_a asociados a las variables.

2.3.3. Reducción de dimensión mediante *Principal Component Analysis*

Principal Component Analysis es una técnica de reducción de dimensionalidad para variables correlacionadas. Esta herramienta apunta a crear un conjunto de variable no correlacionadas de menor dimensión. Esto se logra mediante una transformación lineal que es óptima en términos de capturar la variabilidad del sistema.[29][20] Formalmente, se tiene una matriz X con n observaciones y m variables observadas:

$$X_{m,n} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{pmatrix} \tag{2.7}$$

PCA encuentra una denominada «matriz de carga» $P \in \mathbb{R}^{m \times a}$ en donde $a \leq m$, que relaciona X con los primeros a componentes contenidos en la matriz *score*:

$$T = XP \tag{2.8}$$

Además se tiene que:

1. $Var(t_1) < Var(t_2) < ... < Var(t_a)$
2. $Mean(t_{\mathrm{i}}) = 0, \forall \mathrm{i}$
3. $t_{\mathrm{i}}^T t_k = 0, \forall \mathrm{i} \neq k$
4. No existe ningún otro tipo de expansión ortogonal de a componentes que capture más variabilidad en los datos.

En donde t_{i} corresponde a la i-ésima columna de T. Entonces, esta matriz T representa las proyecciones de la matriz X en el espacio reducido a a dimensiones. Por otro lado, las proyecciones de regreso al espacio de m dimensiones están dadas por:

$$\hat{X} = TP^T \tag{2.9}$$

Además, se tiene que la matriz residual definida como:

$$E = X - \hat{X} \tag{2.10}$$

captura las variaciones generadas por las $m - a$ componentes restantes. Esta matriz tiene teoricamente un ratio *signal-to-noise* bajo por lo que cuando el parámetro a es propiamente escogido, esta matriz representa el ruido.

2.3.4. *Similarity Based Modeling*

Para entender el concepto de *Similarity Based Modeling (SBM)* primero se debe entender los conceptos de métodos no-parámetricos y *Space Kernel Analysis (SKA)*. Por una parte, los métodos no-paramétricos proveen de una herramienta para explicar y/o diagnosticar cuando

se tienen conjuntos de datos complejos. Esto debido a que no dependen de la suposición de que los datos tienen una distribución de probabilidad dada. Por otra parte, los métodos *SKA* son definidos entre un vector y un espacio, a diferencia de los métodos basados en *kernel* que son definidos entre vectores[17]. A continuación, se detallarán los métodos *SKA* para luego especificar un caso de *SKA*: Los modelos basado en similitud.

Un problema típico de aprendizaje de máquinas tiene un vector de entrada $X \in \mathbb{R}^m$ y un vector de salidas $Y \in \mathbb{R}^p$. Además, tiene un conjunto de entrenamiento formado por L pares observados $(X_1, Y_1), ..., (X_L, Y_L)$. El modelo entonces se puede escribir como:

$$Y_{\mathrm{i}} = f(X_{\mathrm{i}}) + \varepsilon_{\mathrm{i}} \tag{2.11}$$

En donde la función $f : \mathbb{R}^m \to \mathbb{R}^p$ es desconocida y se quiere estimar. Tipicamente, se elige una función tal que minimice alguna función de pérdida. Luego se define una función *kernel* $K(X_1, X_2)$ que debe ser positiva, acotada, integrable, con valores en $\mathbb{R}$, simétrica y monótona decreciente con $\|X_1 - X_2\|$. Usualmente y sin pérdida de generalidad, se puede decir que el máximo de $K(X_1, X_2)$ es 1 y se cumple cuando $X_1 = X_2$. Además, una matriz *space kernel* $A \in \mathbb{R}^{L \times L}$ es definida en el conjunto de entrenamiento. Denotando $X_{tr} = [X_1, ..., X_L]$ e $Y_{tr} = [Y_1, ..., Y_L]$, se tiene que la salida Y_n de un *SKA* dada una entrada X_n es:

$$Y_n = \hat{f_{SKA}}(X_n) = \frac{\sum_{\mathrm{i}=1}^{L} Y_{\mathrm{i}} \sum_{j=1}^{L} A_{\mathrm{i}j} K(X_j, X_n)}{\sum_{\mathrm{i}=1}^{L} \sum_{j=1}^{L} A_{\mathrm{i}j} K(X_j, X_n)} \tag{2.12}$$

Reescrito es:

$$Y_n = \frac{Y_{tr} \cdot A \cdot (X_{tr}^T \otimes X_n)}{\mathbf{1} \cdot A \cdot (X_{tr}^T \otimes X_n)} \tag{2.13}$$

en donde $\otimes$ es el operador de similitud y satisface:

$$X_{tr}^T \otimes X_n = [K(X_1, X_n, ..., K(X_L, X_n)]^T \in \mathbb{R}^L \tag{2.14}$$

Por último, los modelos basado en similitud es un caso de *SKA* en donde la matriz A corresponde a $A = (X_{tr}^T \otimes X_{tr})^{-1}$ [17] con

$$X_{tr}^T \otimes X_{tr} = \begin{pmatrix} K(X_1, X_1) & K(X_1, X_2) & ... & K(X_1, X_L) \\ K(X_2, X_1) & K(X_2, X_2) & ... & K(X_2, X_L) \\ \vdots & \vdots & \ddots & \vdots \\ K(X_L, X_1) & K(X_L, X_2) & ... & K(X_L, X_L) \end{pmatrix} \tag{2.15}$$

2.3.5. Matriz de confusión, *Recall* y *Precision*

La matriz de confusión, *Recall* y *Precision* corresponden a formas de evaluar el desempeño de un clasificador. En el caso de dos hipótesis (o clases) H_0 y H_1 se tiene cierta equivalencia. Entonces, la matriz de confusión se representaría como:

Cabe destacar que mientras la matriz sea más parecida a una diagonal, mejor es el desempeño del clasificador. Por otro lado, los términos *Recall* y *Precision* se definen como:

$$Recall = \frac{TP}{TP + FN} \tag{2.16}$$

Tabla 2.1: *Matriz de confusión*

Hecho/Predicción	H_0	H_1
H_0	Verdadero Negativo (TN)	Falso Positivo (FP)
H_1	Falso Negativo (FN)	Verdadero Positivo (TP)

$$Precision = \frac{TP}{TP + FP} \tag{2.17}$$

En donde, los términos TP, FN y FP están definidos en la Tabla 2.1. Entonces *recall* puede interpretarse como la capacidad del clasificador para identificar las observaciones que cumplen la hipótesis H_1. Por otro lado, *precision* corresponde a la tasa de predicho H_1 que efectivamente son H_1.

Capítulo 3

Metodología

Como se muestra en la Figura 3.1 el proceso completo consta de un pre-procesamiento y de dos implementaciones de *SBM*: uno con la información de U-cursos y otro sin estos datos. En primer lugar, el pre-procesamiento consta de la recolección de datos de distintos tipos, para transformarlos en variables. En este sentido, se forma una matriz de observaciones, en donde cada estudiante es representado por una observación con sus respectivas variables. Además, estos datos se separan en conjuntos de entrenamiento y validación. En segundo lugar, se implementan los métodos de *PLS* y *PCA* para seleccionar variables en los dos modelos propuestos. Se hace de esta forma ya que los conjuntos de entrenamiento son diferentes para cada modelo (el modelo sin los datos de U-cursos tiene menos variables). Luego, se pone en funcionamiento *SBM* como se expone en Subsección 2.3.4 con la finalidad de estimar la nota final de los estudiantes del conjunto de validación. Por último, en la etapa de post-procesamiento se identifica si la nota final (predicha y real) corresponde a una reprobación y en base a eso se calcula la matriz de confusión, el *recall* y la precisión.

Dentro del pre-procesamiento la primera etapa es identificar la muestra. Es decir, determinar los estudiantes (con su identificación) que serán el sujeto de estudio. El paso siguiente consiste en la recolección de notas, de antecedentes, de fechas de controles y de *logs* de U-cursos. En primer lugar, las notas recolectadas corresponden a los resultados oficiales de las evaluaciones como controles y examenes. En segundo lugar, los antecedentes consisten en datos sociodemográficos sumado a la información de admisión. Por ejemplo, dentro de los antecedentes están el género del estudiante o su puntaje PSU. Luego, se recogen las fechas de las evaluaciones durante el primer año. Estas evaluaciones consisten en su mayoría de controles, sin embargo también se consideran los controles recuperativos y los examenes. Por último, se reunen los registros (*logs*) de la plataforma U-cursos. Esta plataforma escribe una línea cada vez que un usuario o administrador realiza alguna acción. Para entender cómo se salvan estas acciones se introduce el término «Módulo» y se hace una diferencia entre «leer» un mensaje y «mirar» un módulo. En efecto, Módulo hace referencia a las herramientas de U-cursos. Por ejemplo, la parte de la plataforma que corresponde a material docente es un módulo, también los foros son módulos. Además, la diferencia entre leer un mensaje y mirar un módulo se explica con un ejemplo: cuando hay un mensaje nuevo en el foro, todos los estudiantes del curso son informados de este nuevo elemento. Entonces, el estudiante puede bien ingresar al módulo foro a leer (y comentar el mensaje), o bien ingresar al módulo foro

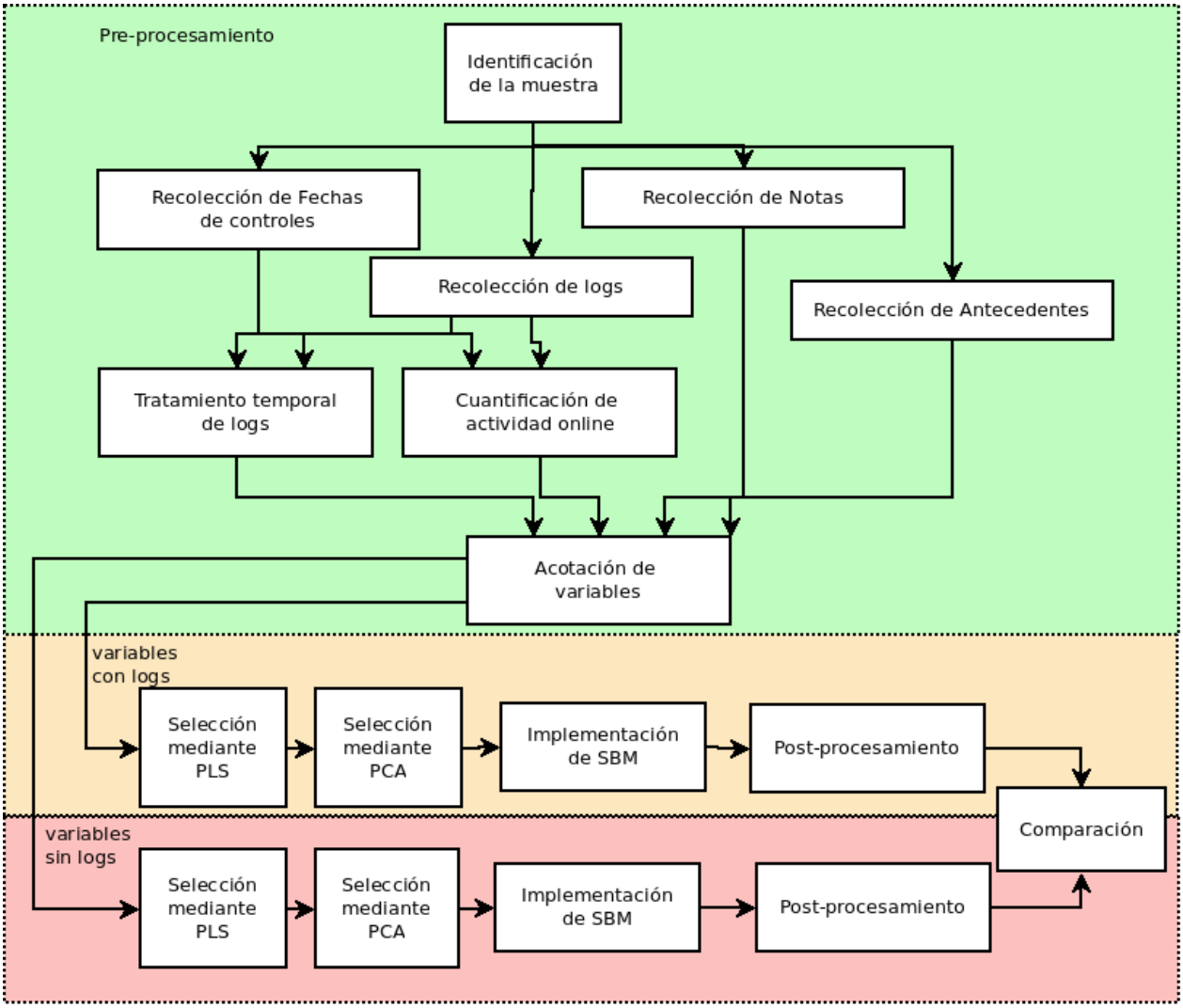

Figura 3.1: *Esquema de la metodología.*

y notar que el mensaje no es de su interés y cerrar sesión. En consecuencia, el servidor de U-cursos registra actividades diferentes en los dos casos, identificando también el módulo.

Luego, toda la información recolectada debe traducirse en variables. En los casos de las notas, información socio-demográfica y de admisión, los datos se consideran directamente como variables. Por otro lado, surge una pregunta importante respecto al cómo llevar los registros de U-cursos a variables que puedan ser implementadas en el modelo matemático. Un acercamiento utilizado en varias situaciones es contar el total de *logins* en la plataforma[21][6][27]. Sin embargo, dentro de los datos disponibles no se identifican los *logins*, sino las acciones realizadas. Si bien se podría determinar un equivalente al inicio de sesión, se considera que se pierde información respecto a las diferentes naturalezas de acciones que puede realizar el estudiante, por ejemplo descargar material docente o mirar que alguien subió un material. En consecuencia, se considera un conteo de acciones separadas por módulo y por operación[1].

[1]El tipo de operación hace una distinción dentro de lo que se puede hacer en un módulo. Por ejemplo, en el módulo foro se hace la diferencia entre publicar, leer y responder un *post*

Además, se tienen en cuenta dos formas de considerar la dimensión temporal. Por un lado se hace una separación por períodos y por otro lado se aplica una medida que da indicios de la distribución de las acciones en el tiempo. En primer lugar, los períodos establecidos para separar los registros de U-cursos se realizan considerando las fechas de los controles y el inicio de clases. En otras palabras, se considera un período como el lapso entre una evaluación y la siguiente, tomando en cuenta que el primer período es entre el inicio de clases y la primera evaluación. Se considera esta forma de ventanas de tiempo ya que los supuestos hechos en la Sección 2.2 se traducen en un comportamiento diferente respecto al período anterior a un control. En otras palabras, se espera que un estudiante con enfoque profundo tenga un número elevado de actividad *online* a lo largo de la ventana de tiempo. Además, se supone que un alumno con una aproximación estratégica tenga primero un número bajo de actividad y vaya aumentando a medida que se acerca el semestre. Por último, se concibe que un estudiante con una orientación superficial tenga un número bajo de actividad concentrada al final de la ventana de tiempo.

Entonces, se crean las variables «cuantitativas», llamadas así en lo que sigue de la Memoria de Título, que consisten en la cuenta de líneas de los registros en un determinado Período, separadas por módulo y por operación. Luego, se plantean las variables «temporales» con el objetivo de indicar una medida de cómo el estudiante ha distribuido el tiempo durante el período entre evaluaciones. En un principio, se pensó utilizar el centroide como medida, tal como se muestra en la Figura 3.2. Entonces, un centroide con una coordenada en el tiempo más cercana a 1 (considerando 0 como la fecha de la evaluación anterior o el inicio de semestre según corresponda) significa que el estudiante ha consultado U-cursos con poca anticipación, en cambio un valor cercano a 0 indica que el estudiante ha consultado U-cursos con más anticipación. Además, se evalúa la opción de utilizar el promedio de estas fechas.

En efecto, al observar la Figura 3.2 se puede ver que el promedio de las fechas es más cercana al momento en donde hay mayor cantidad de actividades. Esto se debe a que el centroide otorga un peso mayor a las actividades más cercanas a 1 (evaluación), puesto que tienen un mayor numero de actividades acumuladas. Entonces, se considera como una mejor opción el hecho de que todas las actividades tengan un mismo peso. En otras palabras, se utiliza el promedio de las fechas de los registros como variable.

Luego todas las variables se acotan utilizando la Ecuación 4.1 para las variables temporales y la Ecuación 4.2 para el resto de variables (ver próximo capítulo).

En los siguientes pasos se implementan dos modelos diferentes, dado que los datos tienen naturalezas distintas. Sin embargo el procedimiento coincide para estos dos modelamientos. Este método consiste en realizar una selección de atributos, luego una predicción y por ultimo un post-procesamiento. En primer lugar, se establecen algoritmos de *feature extracion* para evitar variables que no aportan a las predicciones del modelo. En esta parte se utiliza *PLS* para descartar variables y *PCA* para reducir la dimensionalidad del problema. Además se recuperan las variables descartadas y aquellas que más aportan a la predicción. En segundo lugar, el modelo *SBM* se utiliza para realizar una predicción de la nota final de los estudiantes. Por último, según esta estimación se calcula si el estudiante aprueba o reprueba ($nota_{final} < 4$) y se compara con la situación real. En base a esto se calculan: la matriz de confusión, luego el*recall* y la precisión, para evaluar el desempeño del modelo. Cabe destacar que el *recall*

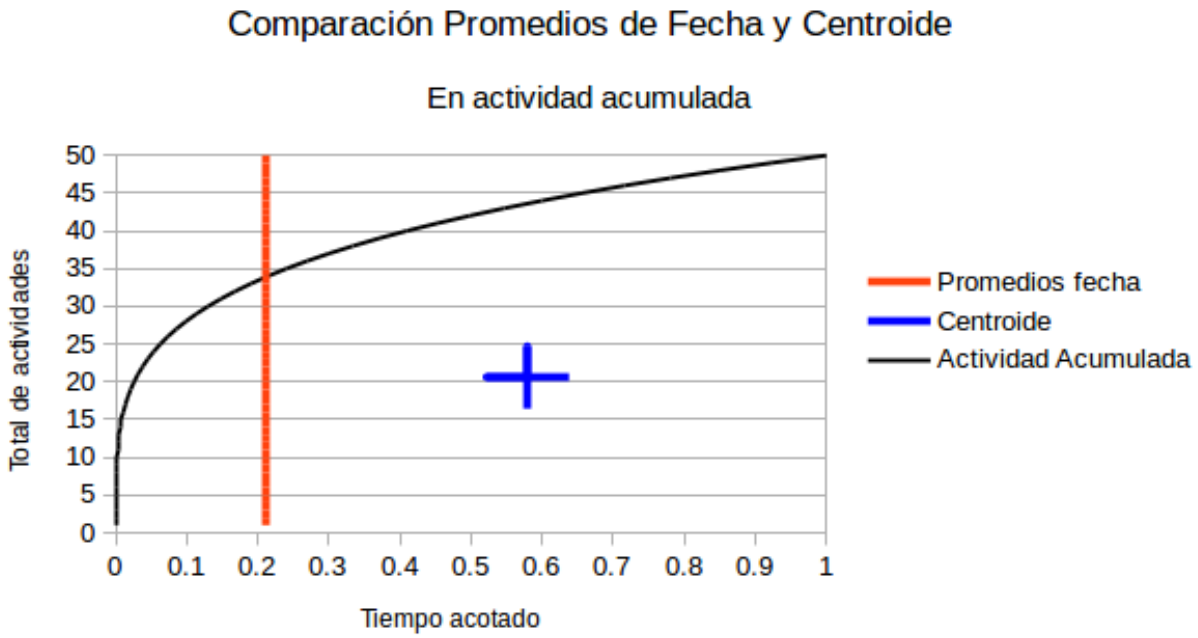

Figura 3.2: *Comparación entre dos formas de obtener una dimensión temporal a partir de la actividad en línea entre 2 evaluaciones. En el eje X, se representa el tiempo entre dos evaluaciones. El valor 1 representa la evaluación hasta donde se ha considerado la información, mientras que el valor 0 representa la evaluación anterior, o en caso de que corresponda, el inicio del semestre. En el eje Y se observa el total absoluto de actividades. Luego, la línea negra es la actividad online acumulada por un estudiante particular. La linea roja indica el promedio de las fechas de los registros de la actividad en línea y por último, la marca azul indica la posición del centroide del área bajo la curva de la actividad acumulada por el estudiante.*

se interpreta como el porcentaje de estudiantes identificados como reprobados dentro de todos los estudiantes que no aprueban. En cambio, la precisión se traduce en el porcentaje de estudiantes que efectivamente reprueban dentro de los cuales el sistema predijo como reprobados. En consecuencia, se considera que el *recall* es más importante dado que el costo de no identificar a un estudiante que repueba es más alto que alarmar a uno que no reprobaría.

Por añadidura, este proceso se repite seleccionando la información por períodos. Por ejemplo, se realiza la predicción considerando las variables que se han podido generar desde el inicio de clases hasta la fecha del control 3 y luego se realiza nuevamente con la misma información más los datos hasta el control 4.

Capítulo 4

Implementación de Modelos Basados en Similitud

4.1. Base de Datos

La base de datos consiste en tres partes: Las notas, los antecedentes y los *logs* generados por U-cursos. Primero, las notas se encuentran en planillas .ods (Formato de LibreOffice Calc), separadas por curso y sección. Segundo, los antecedentes se encuentran en una planilla excel. Estas dos planillas se llevan a .csv (*comma separated value*) para que se puedan trabajar en Python. Tercero, los *logs* de U-cursos se encuentran en una base de datos *MySQL*. Estas últimas tablas se trabajan utilizando Python como interfaz en MySQL. Primero se identifica la muestra del estudio, es decir, se filtran las ID de los estudiantes que ingresaron entre el 2013 y el 2014 a la Facultad dentro de una tabla con la información de estudiantes ingresados desde el 2010 al 2014. Luego, se procede a juntar sus notas, antecedentes e información de U-cursos.

Dentro de las variables en notas que se recolectan están: Notas de controles (6 para los cursos de matemáticas de primer semestre más una nota de control recuperativo), nota de examen y por último la nota final de aprobación. Por otro lado, existen secciones y cursos con información menos detallada.

Sumado a las notas, se encuentran las variables ligadas a los cohortes. Estas variables son por ejemplo: el género o el puntaje PSU. En la Tabla 4.1 se puede observar un detalle de las variables que se toman en cuenta. Además los eventos generados por U-cursos que se utilizan se dividen en módulos y operaciones. En la Tabla 4.2 se observa una descripción de la información utilizada.

Por último, se encuentran las notas de los cursos que se obtienen a partir de las actas.[1] En el caso de los cursos de matemáticas de primer año en primer semestre las actas tienen

[1]Las actas corresponden a las notas oficiales del estudiante en la universidad, estas actas se encuentran en su mayoría estandarizadas. Sin embargo hay casos en los que se omite información respecto al desarrollo del estudiante a lo largo del año.

Tabla 4.1: *Descripción de las variables correspondientes a los antecedentes de las cohortes 2013 y 2014.*

Nombre	$\bar{X}$	σ	Mín-Máx	Descripción
n_intentos	1.01	0.09	1-2	Indica la cantidad de intentos de la persona para ingresar a la facultad.
genero	1.24	0.43	1-2	Indica el género del estudiante.
región	-	-	1-15	Corresponde a la región de procedencia del estudiante. En donde se utiliza la codificación dada por la Subsecretaría de Desarrollo Regional y Administrativo[11].
dependencia	1.81	0.81	1-3	Indica el tipo de establecimiento de procedencia. 3: Municipal; 2: Subvencionado; 1: Particular.
ano_egreso	-	-	2000-2013	Corresponde al año en que el estudiante ha egresado de su establecimiento de educación media.
nem	6.48	0.24	5.3-7	Indica la nota obtenida en la educación media.
psu_x	743.97	29.06	650-840.1	Corresponde al puntaje PSU obtenido. En este caso x puede representar matemáticas, lenguaje, ciencias, historia o el puntaje ponderado para ingresar a la facultad. Se muestran los valores del puntaje PSU ponderado
ranking_em	6.52	13.40	1-79	Ranking en el establecimiento de procedencia.

Tabla 4.2: *Descripción de la información utilizada que ha sido generada por la plataforma u-cursos. La operación 0 corresponde a solo ver el módulo.*

Módulo	Operaciones	Descripción
FORO	0,1,2	1: Leer mensaje, 2: publicar mensaje
MATERIAL_DOCENTE	0, 3	3: descargar material. Las operaciones 1 y 2 son exclusivas para el profesor.
MATERIAL_ALUMNOS	0, 1, 3	1: agregar material, 3: descargar material
MENSAJES	0	En este caso la operacion 0 coincide con escribir mensaje

la información de 7 controles, sumado a esto tienen las notas de examen, de un control recuperativo y por último la nota final del curso. Por otro lado se tienen las fechas en la que los controlores tuvieron lugar. Esta información ha sido recolectada a través de los calendarios proporcionados por la Escuela[2] y por la plataforma Ucampus[3].

Un descubrimiento interesante al observar las variables es la relación entre el número de mensajes escritos en el foro y el género. En la tabla 4.3 se puede observar que en general, los hombres publican mensajes más que las mujeres, de hecho el único caso en que el promedio de mensajes publicados por mujeres es mayor que el de hombres es para el período correspondiente al control recuperativo.

Tabla 4.3: *Promedio del número publicaciones en el módulo FORO por género. El término h«número» indica el período del control en que se cuenta el número de publicaciones. Por ejemplo, FORO_ cu_ op_ 2_ h3 con valores 0.20 para el género Femenino y 0.26 para el masculino indica que entre el control 2 y 3, las mujeres publicaron en promedio 0.2 mensajes y los hombres 0.26.*

Variable	Femenino	Masculino
FORO_cu_op_2_h1	0.13	0.23
FORO_cu_op_2_h2	0.11	0.20
FORO_cu_op_2_h3	0.20	0.26
FORO_cu_op_2_h4	0.10	0.11
FORO_cu_op_2_h5	0.10	0.08
FORO_cu_op_2_h6	0.08	0.14
FORO_cu_op_2_h7	0.16	0.31
FORO_cu_op_2_h8	0.15	0.32

4.2. Pre-procesamiento

4.2.1. *Logs* de U-cursos

La información generada por U-cursos es muy grande, por ejemplo, la base de datos que contiene los datos del 2013 contiene 1568 tablas en donde algunas tablas pueden llegar a tener mas de 2.000.000 de líneas. Esto implica que debe ser pre-procesada para llevar esta información a variables que puedan ser tratadas por el modelo. En este sentido, se divide la información por temporalidad. Esta temporalidad es determinada por las fechas (y hora) de los controles. Es decir se separan las actividades en conjuntos diferentes que se encuentran entre controles, por ejemplo se tiene un conjunto para las actividades desde el inicio del semestre hasta en control 1, luego otro conjunto para las actividades desde el control 1 hasta el control 2 y así sucesivamente.

[2]La Escuela de Ingeniería y Ciencias es responsable por la administración central de los planes de estudio de pregrado, los cursos transversales de formación integral y la coordinación de la enseñanza que se imparte en los departamentos académicos de la Facultad.

[3]Es una plataforma utilizada por la FCFM, en donde se publican informaciones oficiales tales como horarios de evaluaciones y notas, entre otros.

Luego se tienen dos maneras de aproximarse a los datos: Una forma cuantitativa y una forma temporal. En primer lugar, la forma cuantitativa consiste en el total de actividad realizada por el estudiante hasta un determinado control, es decir, se genera una variable que representa el total de actividades en un módulo, considerando una operación y el período en que ocurrieron estas actividades considerando el control como última fecha. En segundo lugar, se considera una forma temporal, que se traduce en el promedio de las fechas en que las actividades de un módulo y operación fueron realizadas. Una muestra de estas dos variables se pueden ver en la Figura 4.1.

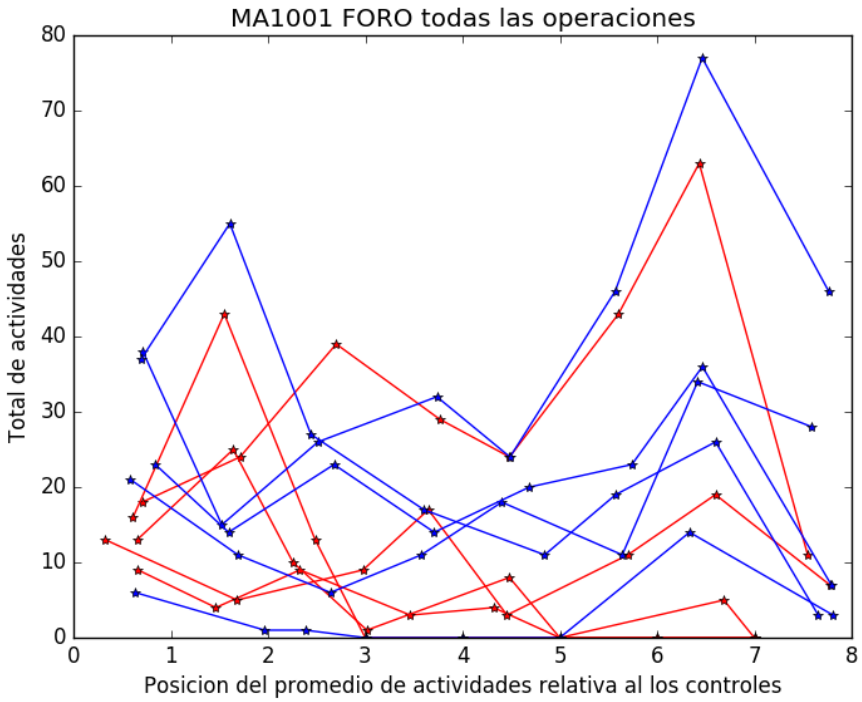

Figura 4.1: *Ejemplo de las variables de los promedios de la actividad en el modulo FORO del curso Introducción al Cálculo (MA1001) contando todas las operaciones. Cada línea representa a uno de los 10 estudiantes elegidos al azar. Además, los segmentos azules corresponden a estudiantes que han aprobado y las líneas rojas a estudiantes que reprobaron.*

En la Figura 4.2 se puede ver un ejemplo de las variables creadas a partir de los *logs*. En estos histogramas se pueden ver los dos tipos de variables, que a simple vista tienen distribuciones diferentes. Por un lado, en las variables temporales se identifican 2 grupos dentro de un mismo histograma. Esto se debe a un grupo que prácticamente no tiene actividad y otro que distribuye de una forma similar a una campana. Por otro lado, las variables cuantitativas distribuyen de forma similar a una Γ o una $\chi - cuadrado$.

4.2.2. Acotación de variables

Para poder implementar SBM las variables deben tener magnitudes similares ya que las distancias entre vectores se puede ver afectada por estas diferencias, En particular en este

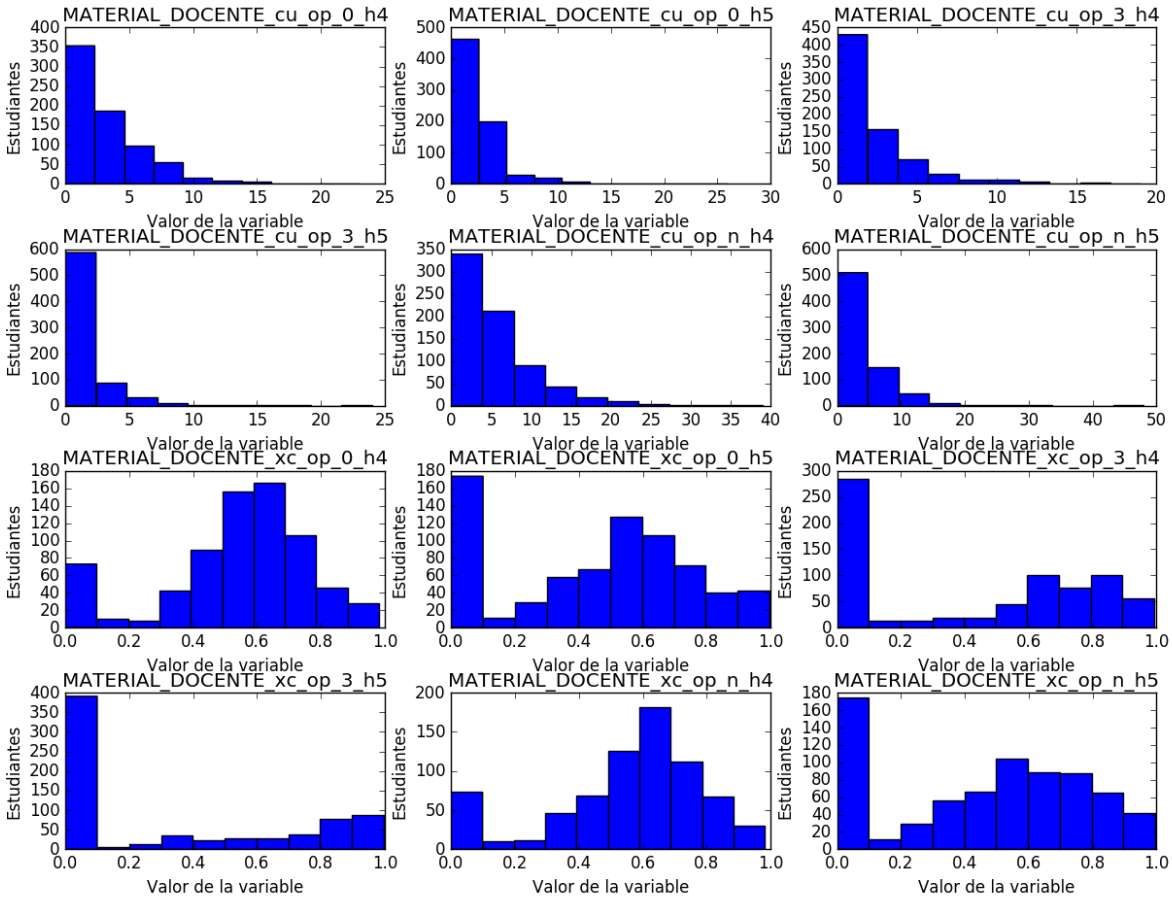

Figura 4.2: *Histograma de las variables relacionadas con los logs del módulo material docente. Las variables que contienen «xc» corresponden a variables temporales y las variables que contienen «cu», corresponden a variables cuantitativas. Luego, se indica la operación, en donde 0 es la operación estándar, 3 corresponde a la descarga de material y n corresponde a la suma de todas. Luego se indica el período en el que se computa la variable, por ejemplo h4 corresponde al período entre el control 3 y el control 4.*

caso en donde el orden de magnitud de variables es demasiado grande. Por ejemplo, el genero es 0 ó 1 y la variable temporal correspondiente a la actividad en el foro hasta el control 1 es del orden de 10^9 dada su resolución en segundos.

Entonces el objetivo de esta parte es que todas las variables tengan valores tales que pertenezcan al intervalo $[0, 1]$. Esto se logra de diferentes formas para distintos tipos de variables: Para las variables temporales se realiza transformación del tiempo en donde el inicio del semestre c_0 o el control anterior $c_{\mathrm{i}-1}$ corresponde a 0 y la información hasta el control i corresponde a 1. Es decir, para cada variable relacionada con el tiempo t_{i} se tiene la transformación de la Ecuación 4.1:

$$t_{\mathrm{i}}^{transformada} = \frac{t_{\mathrm{i}} - c_{\mathrm{i}-1}}{c_{\mathrm{i}} - c_{\mathrm{i}-1}} \tag{4.1}$$

Por otro lado, el resto de las variables v se transforma tomando el mínimo y máximo de cada uno y haciendo una operación similar a lo anterior, esto se puede observar en la Ecuación 4.2:

$$v^{transformada} = \frac{v - \mathrm{mín}}{\mathrm{máx} - \mathrm{mín}} \tag{4.2}$$

4.3. Selección de variables mediante *PLS* y *PCA*

Una vez que las variables han sido llevadas al intervalo $[0, 1]$ se procede a descartar aquellas variables que no aportan información o aquellas que aportan muy poca información. para lograr esto se consideran dos técnicas de *feature extraction*: *PLS* y *PCA*.

4.3.1. Selección de variables con *Partial Least Squares*

En esta sección el objetivo es descartar aquellas variables que no aportan información relevante en el modelo. Para lograr esto, se ajusta el modelo como se explica en la Subsección 2.3.2. Para este ajuste se consideran solamente 2 parámetros relevantes. Puede parecer que utilizar solo dos componentes es una reducción muy grande, sin embargo, como el objetivo no es implementar *PLSR* sino que ver que variables aportan poco se decide utilizar solamente dos componentes. Además, estas dos componentes son las que más información entregan, entonces, agregar otras variables es cada vez menos significativo.

4.3.2. *Principal Component Analysis (PCA)*

En esta sección se explica la implementacion de *PCA*. En primer lugar, se calculan las componentes principales sin eliminar componentes. Luego, se buscan las primeras n componentes tales que la explicabilidad de la varianza sea mayor que 80 %. Esta explicabilidad se

obtiene sumando los valores propios de las primeras n componentes dividido por la suma total de valores propios. Por último, se efectúa la reducción de dimensionalidad considerando las n componentes antes calculadas. En consecuencia, se tienen matrices de entrenamiento y validación más pequeñas, en donde las variables son combinaciones de las variables definidas anteriormente.

4.4. Implementación de *SBM*

Para elegir el kernel se probaron dos kernels diferentes: Triangular y Gaussiano. Para estos dos kernel se debe elegir un parámetro σ en donde se eligió como un tercio del promedio de las distancias entre observaciones. En la Figura 4.3 se observa donde estaría situado el parámetro σ en las distancias observadas. Esto quiere decir que cada vez que se utiliza el operador de similitud entre un vector y la matriz de entrenamiento, se tiene que los vectores que están a un tercio del promedio en cercanía tienen una mayor ponderación.

Una de las razones para utilizar el kernel triangular es la simplicidad en el cálculo matricial, dado que el kernel gaussiano entrega valores muy cercanos a 0 en algunos casos, lo que implica que al invertir una matriz puede causar problemas. Sin embargo, las herramientas computacionales utilizadas no fueron un impedimento para realizar estos cálculos con un kernel gaussiano. Entonces, se decide por utilizar un kernel gaussiano ya que aporta un mejor clasificador que el kernel triangular, debido a que asigna una ponderación casi 0 a las variables más lejanas.

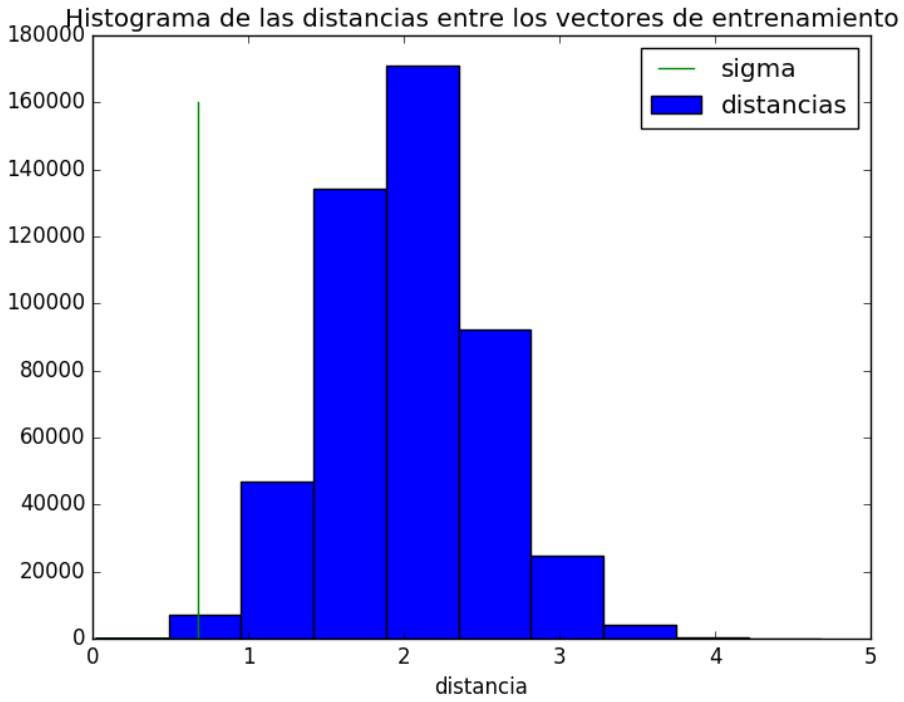

Figura 4.3: *Histograma de las distancias entre observaciones (estudiantes). La línea verde muestra el parámetro* σ.

Capítulo 5

Resultados

En este capítulo se muestran las variables descartadas através de *PLS*, las variables con mayor importancia según *PCA* y los desempeños de los modelos *SBM*. En primer lugar, se muestran las variables proyectadas en los vectores de carga del modelo *PLS*, en este sentido se busca descartar aquellas características que se encuentran cerca del origen, lo que significa que tienen poco peso. En segundo lugar, se exponen las variables con mayor importancia para las componentes de *PCA*. Por último se exponen los resultados de las clasificaciones con los *SBM*

5.1. Pre-procesamiento

5.1.1. Selección de variables con *PLS*

En la Figura 5.1 se observan los pesos de las variables proyectadas en los dos primeros vectores de carga. Cabe destacar que estos dos vectores son los que agregan mayor explicabilidad al problema, por lo tanto se utilizan para analizar las variables a descartar. Entonces, se fija un umbral de 0.05 que es representado por el circulo rojo de la Figura 5.1 que encierra las variables que se descartan en esta parte del proceso. Luego, las variables eliminadas se indican en la Tabla 5.1. En esta tabla se mencionan las variables que se eliminaron en el proceso de *cross-validation* considerando la información hasta el 4 y 5 control.

En primer lugar, las variables eliminadas correspondientes a los antecedentes son la región, el puntaje PSU de lenguaje y la dependencia del colegio (Particular, Subvencionado y Municipal). En segundo lugar, las variables del módulo MENSAJES corresponden a los controles 1 y 5, es decir el primer control y el control recuperativo. En tercer lugar, la variable del módulo MATERIAL DOCENTE eliminada corresponde a la dimensión temporal de la actividad estándar más cercana al inicio de clases. En cuarto lugar, se destaca la presencia de 22 variables correspondientes al módulo MATERIAL ALUMNOS. En último lugar, se eliminan las variables correspondientes al FORO del control 1 y control recuperativo. Por último, se elimina solamente una variable correspondiente al FORO en el control 4.

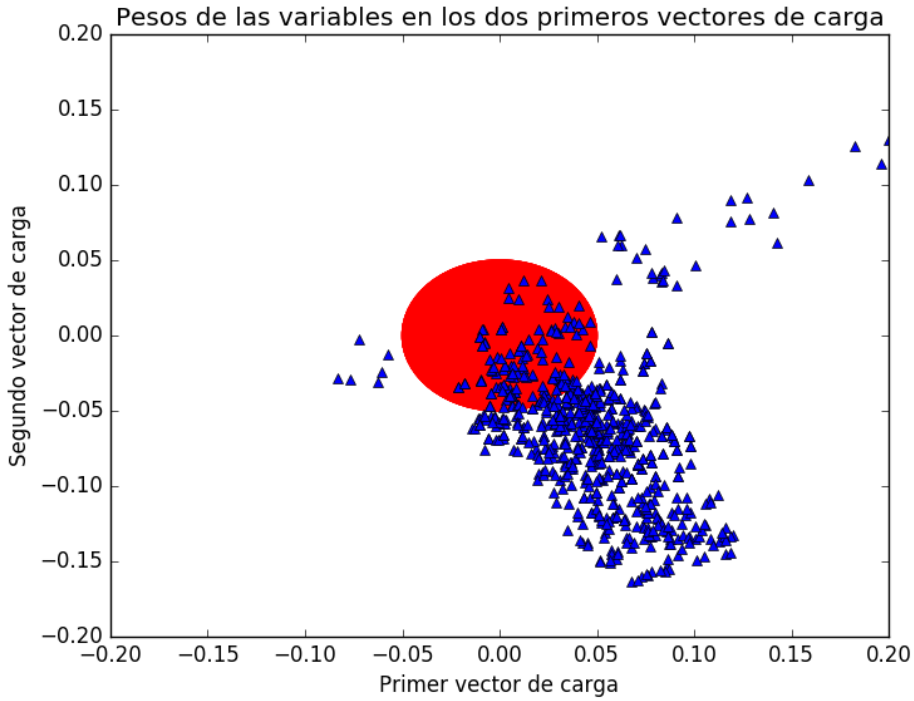

Figura 5.1: *Pesos de variables en los primeros 2 vectores de carga.*

5.1.2. Variables PCA

Para saber qué variables contribuyen más al modelo matemático se identificaron las 10 variables que tienen un mayor coeficiente para calcular la primera o segunda componente. Entonces, por cada una de las 6 iteraciones de *cross-validation* se contaron las veces en que cada variable está dentro de las 10 más importantes. Luego, a esta cuenta se le llama «ocurrencia».

En la Tabla 5.2 se pueden observar las variables que tienen una mayor contribución a la primera componente y en la Tabla 5.3 se pueden observar las variables que contribuyen más a la segunda componente. En efecto, se destacan las variables de antecedentes correspondientes al año de egreso de enseñanza media, el promedio de notas de enseñanza media, la PSU de ciencias, la PSU de matemáticas, la ponderación obtenida al ingresar a la Facultad, el género y el ranking de enseñanza media. Por otro lado, dentro de las variables relacionadas con la plataforma *LMS* se destacan: número de publicaciones en el FORO antes del primer y tercer control.

En la segunda componente, se ubican solamente variables relacionadas con U-cursos. Además se destaca el promedio de las fechas de operaciones estándar y totales antes del primer control en el módulo MATERIAL ALUMNOS. También, se considera el promedio de las fechas de las descargas y suma de todas las operaciones antes del primer control en el módulo MATERIAL DOCENTE.

5.2. Resultados de la clasificación con *SBM*

La variable en la que consiste la clasificación es una variable binaria que indica si el estudiante aprueba o reprueba el curso. Se considera esta variable ya que es un indicador de rendimiento académico que permite efectuar una clasificación en base a los requerimientos mínimos que exige la Facultad para el avance en los cursos.

En la Figura 5.2 se observa una comparación entre clasificadores, en términos de *recall*. Por una parte, el clasificador que incluye la información de U-cursos (línea azul) obtiene un rendimiento de 64,75 % al inicio. Luego, el clasificador mejora monótonamente al incluir más información, llegando a un 81,9 % en el 5 control. Sin embargo, se observa un estancamiento en el momento del control recuperativo. Finalmente, el clasificador término con un *recall* de 86,72 %.

Por otra parte, el clasificador que no utiliza los registros inicia con un valor cercano a 50 %, es decir al valor que se obtendría con una moneda. Luego, este sistema empieza a tener una mejora monótona desde el cuarto control, logrando un 97,56 % al momento del examen. Sin embargo, teniendo la información del examen se puede saber de forma determinante si los estudiantes reprobarán.

En efecto, ya al segundo control el clasificador con U-cursos supera al otro en un 13,47 %. Luego, esta diferencia se mantiene cercana al 15 % hasta el quinto control. También, la mayor diferencia se observa en el cuarto control, con un 22,71 %. Sin embargo, el clasificador que no tiene la información del *LMS* vence a partir del sexto control con un 0,94 % y termina con una diferencia de 10,98 %.

La comparación entre clasificadores desde el punto de vista de la precisión se puede observar en la Figura 5.3. Por una parte, el clasificador que considera la información de U-cursos inicia con un rendimiento del 41,58 %. Luego aumente monótonamente hasta el cuarto control. A partir de ese momento, y durante el control recuperativo y el quinto control, hay un estancamiento en terminos de precisión a valores cercanos al 60 %. Luego, vuelve a mejorar con la información del control 6 y del examen.

Por otra parte, el clasificador sin la información de U-cursos inicia su desempeño con un 24,19 %. Este valor es cercano al que se tendría al identificar a los reprobados lanzando una moneda. Después, este controlador mejora lentamenta hasta un 35,78 en el quinto control. Luego, hay un cambio de pendiente abrupto hasta llegar a un 64,86 % .

Entonces inicialmente, hay una diferencia de 17,39 %, esta diferencia se mantiene en el orden de 15 %-20 % exceptuando el cuarto control en donde se obtiene la máxima de la direfencia con un 27,64 %. Luego, en el control 6 y examen esta diferencia se hace menos significativa, terminando con un 7,66 %.

En conclusión, se pueden identificar dos períodos largos: el primero desde el control 2 hasta el control 5 y el segundo cosiderando el control 6 y examen. En el primero se observa una superioridad notable del clasificador con U-cursos, superando en términos de *recall* y precisión en valores cercanos al 15 %. Luego en el segundo período esta diferencia es menor

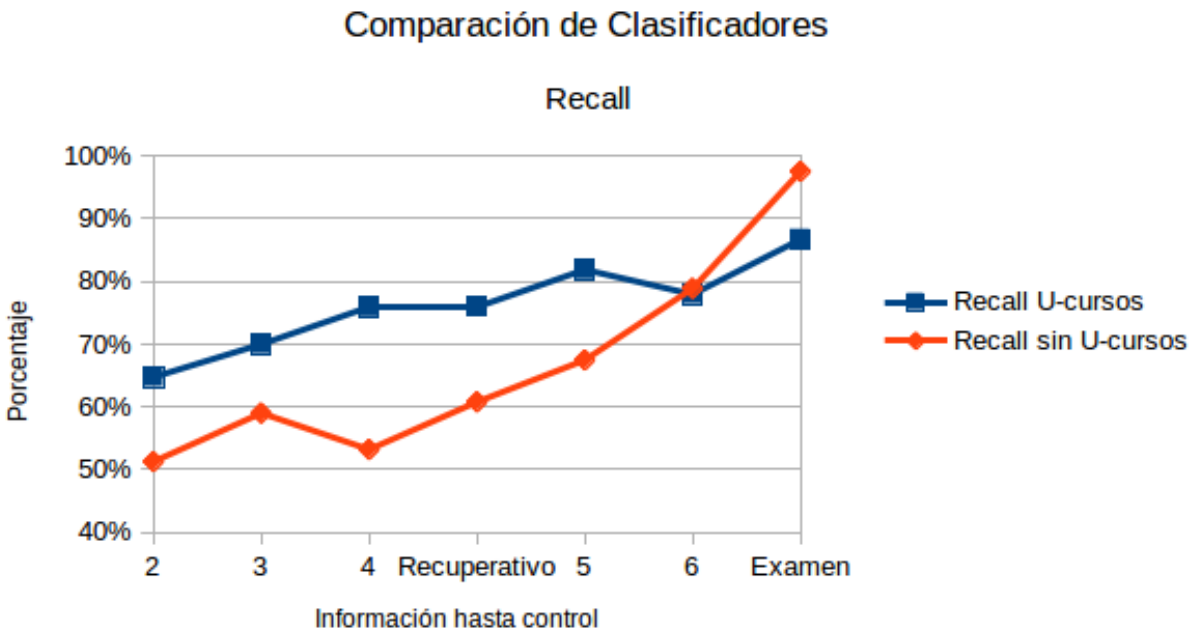

Figura 5.2: *Comparación de clasificadores en términos de recall al ejecutarlo en distintos momentos del semestre. En el eje X se observan las evaluaciones del curso Introducción al cálculo, en orden cronológico y exceptuando el control 1, debido a la poca actividad online de las primeras semanas. Entonces, este eje indica la cantidad de información que se ha utilizado, p. ej. el numero 3 significa que se incluyen las notas (y actividad en línea según corresponda) de los controles 1, 2 y 3. Luego, el eje Y muestra el porcentaje del recall obtenido. Por un lado, la línea azul representa los resultados obtenidos al ejecutar el clasificar con toda la información: antecedentes socio-demográficos y de admisión, notas y variables relacionadas con los logs de U-cursos. En cambio, la línea naranja muestra el desempeño del clasificador al no incluir la información de la plataforma LMS, manteniendo las notas de controles y antecedentes socio-demográficos y de admisión.*

e incluso, el clasificador sin U-cursos supera al otro clasificador.

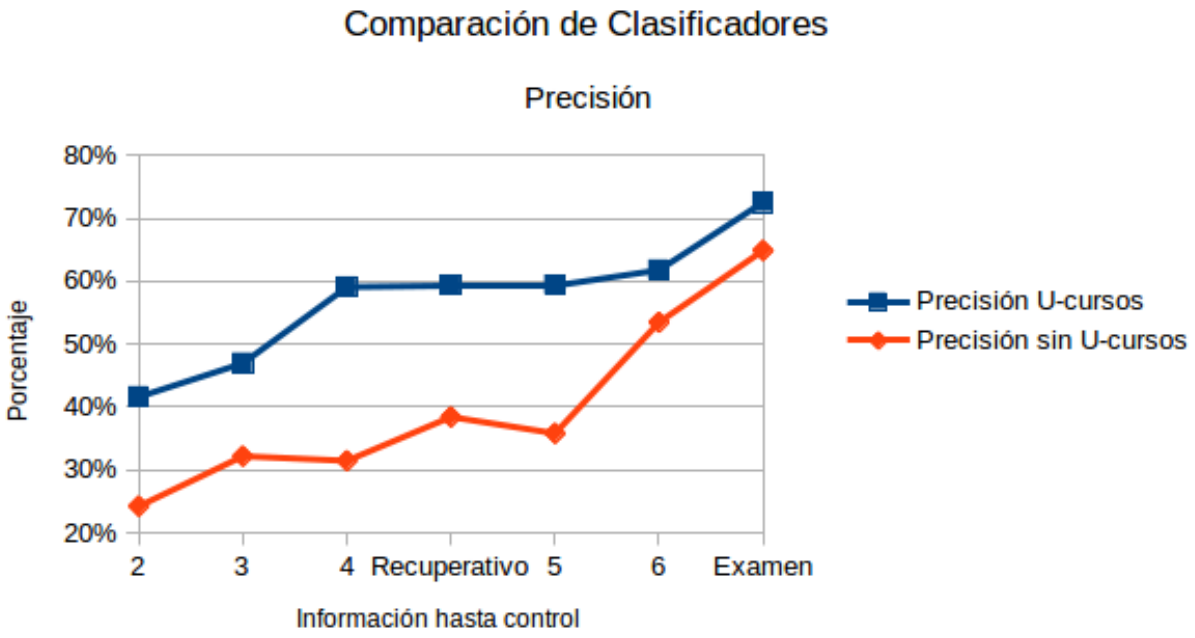

Figura 5.3: *Comparación de clasificadores en términos de precisión al ejecutarlo en distintos momentos del semestre. En el eje X se observan las evaluaciones del curso Introducción al cálculo, en orden cronológico y exceptuando el control 1, debido a la poca actividad online de las primeras semanas. Entonces, este eje indica la cantidad de información que se ha utilizado, p. ej. el numero 3 significa que se incluyen las notas (y actividad en línea según corresponda) de los controles 1, 2 y 3. Luego, el eje Y muestra el porcentaje de la precisión obtenida. Por un lado, la línea azul representa los resultados obtenidos al ejecutar el clasificar con toda la información: antecedentes socio-demográficos y de admisión, notas y variables relacionadas con los logs de U-cursos. En cambio, la línea naranja muestra el desempeño del clasificador al no incluir la información de la plataforma LMS, manteniendo las notas de controles y antecedentes socio-demográficos.*

Tabla 5.1: *Variables eliminadas durante PLS. Las variables cuantitativas corresponden a la suma de actividades de la operación descrita durante el período de control que se menciona. En cambio, las variables temporales corresponden al promedio de las fechas en donde ocurrieron estas actividades del período indicado. Además, el período de control corresponde al lapso entre el control indicado y el control anterior, o en el caso del control 1 corresponde al inicio de clases.*

Módulo	tipo de variable	operación	período de control
Region			
PSU lenguaje			
Dependencia			
MENSAJES	Temporal	Todas	Control 5
MENSAJES	Temporal	Todas	Control 1
MENSAJES	Temporal	Estándar	Control 5
MENSAJES	Temporal	Estándar	Control 1
MATERIAL DOCENTE	Temporal	Estándar	Control 1
MATERIAL ALUMNOS	Temporal	Todas	Control 4
MATERIAL ALUMNOS	Temporal	Todas	Control 2
MATERIAL ALUMNOS	Temporal	Bajar material	Control 5
MATERIAL ALUMNOS	Temporal	Bajar material	Control 4
MATERIAL ALUMNOS	Temporal	Bajar material	Control 3
MATERIAL ALUMNOS	Temporal	Bajar material	Control 2
MATERIAL ALUMNOS	Temporal	Agregar material	Control 5
MATERIAL ALUMNOS	Temporal	Agregar material	Control 4
MATERIAL ALUMNOS	Temporal	Agregar material	Control 2
MATERIAL ALUMNOS	Temporal	Estándar	Control 4
MATERIAL ALUMNOS	Temporal	Estándar	Control 2
MATERIAL ALUMNOS	Cuantitativa	Todas	Control 4
MATERIAL ALUMNOS	Cuantitativa	Todas	Control 3
MATERIAL ALUMNOS	Cuantitativa	Bajar material	Control 5
MATERIAL ALUMNOS	Cuantitativa	Bajar material	Control 4
MATERIAL ALUMNOS	Cuantitativa	Bajar material	Control 3
MATERIAL ALUMNOS	Cuantitativa	Bajar material	Control 2
MATERIAL ALUMNOS	Cuantitativa	Agregar material	Control 5
MATERIAL ALUMNOS	Cuantitativa	Agregar material	Control 4
MATERIAL ALUMNOS	Cuantitativa	Agregar material	Control 2
MATERIAL ALUMNOS	Cuantitativa	Estándar	Control 4
MATERIAL ALUMNOS	Cuantitativa	Estándar	Control 3
FORO	Temporal	Todas	Control 1
FORO	Temporal	Escribir mensaje	Control 5
FORO	Temporal	Estándar	Control 1
FORO	Cuantitativa	Escribir mensaje	Control 5
FORO	Cuantitativa	Escribir mensaje	Control 4

Tabla 5.2: *Variables con mayor contribución a la primera componente de PCA. Además, las variables relacionadas con u-cursos tienen el siguiente formato: «Módulo»_«Cuantitativa (cu) o Temporal (xc)»_«Operación»_«Información hasta control». Por ejemplo, MATERIAL_ALUMNOS_cu_op_0_h3 corresponde a la información de material alumnos, en forma cuantitativa (total de actividad), de la operación estándar y contando la información entre el control 2 y el 3. Por otro lado, la operacion «op_n» corresponde al total de actividades en el módulo.*

Variable	ocurrencia
FORO_cu_op_2_h1	4
FORO_cu_op_2_h3	5
MATERIAL_ALUMNOS_cu_op_0_h3	1
MATERIAL_ALUMNOS_cu_op_0_h4	3
MATERIAL_ALUMNOS_cu_op_3_h5	1
MATERIAL_ALUMNOS_cu_op_n_h4	1
ano_em	6
nem	6
psu_cie	6
psu_len	2
psu_mat	6
psu_pond	6
rank_em	6
region	1
sexo	6

Tabla 5.3: *Variables con mayor contribución a la segunda componente de PCA. Además, las variables relacionadas con u-cursos tienen el siguiente formato: «Módulo»_ «Cuantitativa (cu) o Temporal (xc)»_ «Operación»_ «Información hasta control». Por ejemplo, MATERIAL_ALUMNOS_cu_op_0_h3 corresponde a la información de material alumnos, en forma cuantitativa (total de actividad), de la operación estándar y contando la información entre el control 2 y el 3. Por otro lado, la operacion «op_n» corresponde al total de actividades en el módulo.*

Variable	ocurrencia
FORO_xc_op_0_h1	1
FORO_xc_op_0_h2	1
FORO_xc_op_1_h1	2
FORO_xc_op_1_h2	1
FORO_xc_op_n_h1	1
FORO_xc_op_n_h2	1
MATERIAL_ALUMNOS_cu_op_0_h1	1
MATERIAL_ALUMNOS_cu_op_0_h2	1
MATERIAL_ALUMNOS_cu_op_n_h1	2
MATERIAL_ALUMNOS_cu_op_n_h2	2
MATERIAL_ALUMNOS_xc_op_0_h1	4
MATERIAL_ALUMNOS_xc_op_0_h2	2
MATERIAL_ALUMNOS_xc_op_0_h5	2
MATERIAL_ALUMNOS_xc_op_n_h1	4
MATERIAL_ALUMNOS_xc_op_n_h2	2
MATERIAL_ALUMNOS_xc_op_n_h5	2
MATERIAL_DOCENTE_xc_op_0_h1	2
MATERIAL_DOCENTE_xc_op_0_h2	2
MATERIAL_DOCENTE_xc_op_0_h4	2
MATERIAL_DOCENTE_xc_op_0_h5	2
MATERIAL_DOCENTE_xc_op_3_h1	4
MATERIAL_DOCENTE_xc_op_3_h2	1
MATERIAL_DOCENTE_xc_op_3_h4	2
MATERIAL_DOCENTE_xc_op_3_h5	2
MATERIAL_DOCENTE_xc_op_n_h1	4
MATERIAL_DOCENTE_xc_op_n_h2	2
MATERIAL_DOCENTE_xc_op_n_h4	2
MATERIAL_DOCENTE_xc_op_n_h5	2
MENSAJES_xc_op_0_h5	2
MENSAJES_xc_op_n_h5	2

Capítulo 6

Discusión

En este capítulo se comentan los resultados obtenidos durante el proceso de pre-procesamiento y clasificación del algoritmo. En primer lugar, se comentan las variables descartadas por *PLS*, considerando interpretaciones y posibles mejoras. En segundo lugar, se habla sobre las variables con mayor relevancia considerando *PCA*. En tercer lugar, se tratan las implicaciones de la clasificación. En cuarto lugar, se reflexiona sobre posibles mejoras al modelo.

Por un lado se observa que la PSU de lenguaje, la región de procedencia y el tipo de establecimiento (dependencia) no tienen tanta incidencia. En primer lugar, se puede concluir que el puntaje PSU de lenguaje no se relaciona con las habilidades que se miden en introducción al cálculo. En segundo lugar, la región de procedencia debería tener incidencia, ya que el período de adaptación para estudiantes de región es más difícil. Entonces, esto podría originarse en la forma en que se trató la variable región, es decir, ordenándola de norte a sur y no en base a otro criterio que se relacione con la dificultad de un estudiante para adaptarse en primer año (ej. lejanía con la Universidad, nivel socio-económico región, etc.). Por último, el tipo de establecimiento se trata de una forma similar a la región, con la diferencia de que solo existen 3 posibilidades y no 15. Por lo tanto, se podría tratar como 3 variables binarias distintas o incluir una separación por colegio emblemático.

Luego, se eliminan variables relacionadas con el módulo MENSAJES, MATERIAL DOCENTE (solo control 1) y FORO en los controles 1 y 5, correspondiente al primer control y al control recuperativo. Entonces, se estima que éstos lapsos de tiempo son diferentes a la normalidad del semestre ya que son contemporáneos al inicio de clases y a la semana de controles recuperativos. Esta diferencia se debe a que el inicio de clases en la Facultad tiene muchas actividades de integración, además, en la semana de controles recuperativos no hay clases y los estudiantes pueden decidir no rendir estos controles.

Después, se destaca la presencia de las variables provenientes del módulo de MATERIAL ALUMNOS. Esto significa que para efectuar las predicciones casi no se utiliza la información contenida en los módulos de material alumnos. Sin embargo, los estudiantes que comparten material o bien descargan material que ha sido subido por compañeros tienen un mayor compromiso con la universidad, por ende un interés mayor en estudiar ligado a un enfoque profundo o estratégico, que podría implicar una mayor tendencia a no reprobar. Entonces,

¿Dónde se origina esta aparente contradicción?

Una opción es que durante algunos períodos estas variables tienen una baja actividad, lo que se traduce en que los estudiantes que hicieron algo sean vistos como *outliers* por el algoritmo y no presenten cercanía con otros estudiantes en esta dimensión.

Respecto de las variables con mayor contribución a la primera componente de *PCA* se tienen variables ligadas a los antecedentes que se podrían identificar dos grupos: unas ligadas a rendimientos anteriores (puntaje PSU Matemáticas, Ciencias y ponderado, NEM, Ranking) y otras ligadas a información personal (Género, Año egreso de enseñanza media). En primer lugar, se puede decir que las habilidades medidas en introducción al cálculo se relacionan con el puntaje de Matemáticas y Ciencias. Además, el NEM y el Ranking dan un indicio del enfoque de aprendizaje que tiene el estudiante, suponiendo que un mayor NEM y/o Ranking se relaciona con un aprendizaje profundo o estratégico. En segundo lugar, se tiene que el año de egreso de enseñanza media se relaciona con el tiempo que ha pasado antes de entrar a la facultad, o si el estudiante se ha cambiado de carrera o Universidad. En efecto, los estudiantes que se han cambiado de carrera tienen una menor convicción de seguir estudiando[28][5]. Por otro lado, el porcentaje de mujeres en la Facultad es considerablemente menor al porcentaje de hombres. En consecuencia, esto se traduce en una mayor dificultad para el género femenino en términos de adaptación.

También, en la primera componente de *PCA* destacan las variables de U-cursos del módulo FORO, lo que parece razonable, ya que una mayor participación en los foros implica mayor compromiso con el aprendizaje. En otras palabras, un acercamiento al aprendizaje profundo, en donde el estudiante tiene interés en resolver sus dudas y discutir, además de un compromiso y una motivación mayor.

Por otra parte, en la segunda componente de *PCA* se observa la dimensión temporal de las operaciones estándar y totales en el módulo de MATERIAL ALUMNOS, de las operaciones descargas y totales en el módulo MATERIAL DOCENTE, en el primer control. Al confrontar este resultado con las variables eliminadas en el proceso de *PLS* en donde gran parte de las variables eliminadas correspondían a MATERIAL ALUMNOS. Sin embargo, la variable importante corresponde al primer período, esto se relaciona con estudiantes que inician el semestre con actividad en la plataforma, compartiendo y buscando funcionalidades desde el inicio, lo que da cuenta de un interés académico mayor y también un acercamiento al aprendizaje profundo. A esta confrontación, se suman las variables de MATERIAL DOCENTE, en donde se observa que las variables importantes son las descargas y actividad total, mientras que la menos importante es la operación estándar. Esto se debe a que la operación descarga requiere una proactividad mayor que simplemente observar que hay nuevo material (operación estándar). Por lo tanto, esta pro-actividad se relaciona con el enfoque profundo de Marton y Säljö [22][4].

Por último, se observa una mejora considerable en términos de *recall* y precisión cuando se utilizan los datos de U-cursos. En primera instancia, al incluir los *logs* se obtiene una mejora del 15 %. Esto quiere decir que si como Facultad se quiere tomar una acción respecto de los estudiantes que reprobaren, en una etapa temprana debería utilizarse el modelo con los registros de U-cursos.

También, el modelo utilizado, *SBM*, es un buen modelo ya que se logra realizar una predicción mejor que la que se tendría lanzando una moneda para decidir si los estudiantes aprueban o reprueban. Ésto se cumple en ambos modelamientos y con distintas cantidades de datos. El único caso similar a la predicción con una moneda es cuando se tiene la menor cantidad de datos: El período del control 2 y sin los registros de U-cursos. Además, al ser no-paramétrico no se requiere hacer suposiciones sobre el fenómeno.

Además, el hecho de que la predicción mejore al utilizar los registros de U-cursos, implica que la plataforma *LMS* contiene información, que antes no se tomaba en consideración, sobre los procesos de aprendizaje en los estudiantes.

La implementación de un sistema que permita identificar a estudiantes en peligro de reprobar a nivel de Universidad podría traer beneficios en ámbitos educacionales y económicos. En primer lugar, investigadores que estudian el problema de la reprobación tendrían una herramienta para optimizar sus esfuerzos y centrarse en los estudiantes que posiblemente reprueben. También, se podrían identificar buenas prácticas para el uso del *LMS* y los profesores podrían observar en tiempo real si algunas intervenciones/innovaciones de ellos mejorarán los rendimientos de los estudiantes. Además, la Facultad podría implementar acciones que ayuden a reducir el número de reprobados y por lo tanto reducir el costo de crear nuevas secciones del curso el semestre siguiente. Por otro lado, los estudiantes podrían tener alertas tempranas para pedir ayuda y poder reducir su riesgo de perder becas o pago adicional del arancel, además del costo emocional de reprobar un curso.

Los beneficios de un sistema de esta naturaleza son mayores cuando se implementan a nivel ínter-universidad, ya sea nacional o internacional. Sumado a los bondades anteriores, un sistema de esta naturaleza permitiría caracterizar a estudiantes y realizar comparaciones entre universidades, agregando una característica más al perfil del estudiante. Se crea además una herramienta más para evaluar políticas a nivel directivo, que se podría traducir en un indicador para los Programas de Desarrollo Institucionales de las universidades.

Capítulo 7

Conclusión

La presente memoria de título presenta un modelo empírico que busca predecir la reprobación de estudiantes de ingeniería en primer año en el curso de Introducción al Cálculo. En efecto, el modelo creado relaciona antecedentes socio-demográficos, información de admisión, notas y registros de U-cursos (la plataforma *LMS* utilizada en la FCFM) con la aprobación y reprobación del curso observado. La definición de ventanas de tiempo para el procesamiento de los registros de U-cursos se presenta como una manera innovadora de medir la componente temporal de los datos, ya que se relaciona con las teorías del aprendizaje más aceptadas en el ámbito académico. También, el modelo utilizado(SBM) es nuevo dentro de las aplicaciones de LA, en particular, nuevo en la FCFM. El hecho de que este modelo sea empírico permite reducir el número de supuestos respecto de los procesos de aprendizaje. Además, se ven resultados favorables lo que implica que este modelo es una alternativa a considerar para investigaciones futuras.

En los datos usados se encuentran estudiantes que son en su mayoría hombres, han tenido un elevado puntaje PSU, NEM y un elevado puesto en sus colegios. Además, la cantidad de datos que proporciona U-cursos hace que sea necesario procesarlos antes de que sean usados en un modelo. Por otro lado, la naturaleza diversa de estos datos presenta un desafío en la forma de relacionarlos. Por añadidura, toda esta información es propiedad de la Universidad y permite entender mejor los procesos de aprendizaje de sus estudiantes.

Además, se identifican variables relacionadas con los antecedentes y con el *LMS* que dan indicios de los comportamientos explicados por las teorías de Asikainen [4], Marton y Saljo[22]. En efecto, por el lado de las variables relacionadas con los antecedentes se tienen: el año de egreso de enseñanza media, NEM y el Ranking. Por el lado de las variables en el *LMS* se tienen: variables relacionadas con el FORO, MATERIAL DOCENTE y MATERIAL ALUMNOS. Sin embargo, se identifican variables que no son pertinentes, ya sea por la forma en que se definieron (p. ej. región de procedencia) o por los períodos que no son comunes (p. ej. Semana de controles recuperativos).

Respecto de los resultados obtenidos, el desempeño del clasificador mejora considerablemente(15 %) al incluir la información de U-cursos. Esto implica que los datos proporcionados por el *LMS* contienen información valiosa respecto de los procesos de aprendizaje de los es-

tudiantes. Por lo tanto, se propone que para tomar decisiones respecto de la intervención de estudiantes, focalizar esfuerzos al investigar, evaluar intervenciones/innovaciones se debe incluir la información contenida en U-cursos porque permite identificar de mejor forma a los estudiantes en riesgo de reprobar.

Bibliografía

[1] E. Aguiar, G. A. Ambrose, N. V. Chawla, V. Goodrich, and J. Brockman. Engagement vs performance: Using electronic portfolios to predict first semester engineering student persistence. volume 1, pages 7–33, 2014.

[2] C. Aguirre. Superación académica en primer año de ingeniería y ciencias: Mecanismos de permanencia y mejoramiento académico. *Memoria de ingeniero civil industrial*, 2016. Santiago, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas.

[3] P. Arroway, G. Morgan, M. OḰeefe, and R. Yanosky. *Learning Analytics in Higher Education*. CO: ECAR, Louisville, 2016.

[4] H. Asikainen, A. Parpala, V. Virtanen, and S. Lindblom-Ylänne. The relationship between student learning process, study success and the nature of assessment: A qualitative study. *Studies in Educational Evaluation*, 39(4):211 – 217, 2013.

[5] J. P. Bean. Dropouts and turnover: The synthesis and test of a causal model of student attrition. volume 12, pages 155–187, 1980.

[6] J. Campbell. Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study. *Tesis de Doctorado*, 2007. Indiana, USA, Purdue University.

[7] S. Celis, L. Moreno, P. Poblete, J. Villanueva, and R. Weber. Design and implementation of a learning analytics toolkit for teachers. *Revista Ingeniería de Sistemas*, XXIX:5–24, 2015.

[8] Hanover Research Center. *Predicting College Student Retention*. Hanover Research Center, 2011.

[9] M. A. Chatti, A. L. Dyckhoff, U. S., and H. Thüs. A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5-6):318–331, 2012.

[10] M. M. Chemers, L. Hu, and B. F. Garcia. Academic self-efficacy and first-year college student performance and adjustment. *Journal of Education Psychology*, 93(1):55–64, 2001.

[11] Subsecretaría de Desarrollo Regional y Administrativo. Codigo unico territorial. 2012.

[12] Escuela de Ingeniería y Ciencias Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile. Hechos y cifras. 2016. [en línea] http://escuela.ingenieria.uchile.cl/la-escuela/124152/hechos-y-cifras [consulta: 27 diciembre 2016].

[13] R.O. Duda, P.E. Hart, and D.G. Stork. Introduction. En: Pattern Classification. *New York: John Wiley and Sons*, 33(2):1–19, 2008.

[14] A. L. Dyckhoff, D. Zielke, M. Bültmann, M. A. Chatti, and U. Schroeder. Design and implementation of a learning analytics toolkit for teachers. *Journal of Educational Technology & Society*, 15(3):58–76, 2012.

[15] N. Entwistle. Motivation and approaches to learning: Motivating and conceptions of teaching. In S.A. Brown and Thompson. *Motivating students*, pages 15–23, 1998.

[16] Center for Learning and Performance Technology. Top 200 tools for learning. 2016. [en línea] http://c4lpt.co.uk/top100tools/top-200-tools-for-learning/ [consulta: 14 de Marzo 2017].

[17] L. Gong and D. Schonfeld. Space kernel analysis. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1577–1580, 2009.

[18] IBM. What is big data? 2012. [en línea] http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html [consulta: 27 diciembre 2016].

[19] S. B. Kotsiantis. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37(4):331–344, 2012.

[20] A. León. Detección de anomalías en procesos industriales usando modelos basados en similitud. *Memoria de ingeniero civil electricista*, 2012. Santiago, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas.

[21] L. Macfadyen and S. Dawson. Mining lms data to develop an early warning system for educators: A proof of concept. volume 54, pages 588–599, 2010.

[22] F. MARTON and R. SÄLJÖ. On qualitative differences in learning: I—outcome and process*. *British Journal of Educational Psychology*, 46(1):4–11, 1976.

[23] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118:62 – 69, 2012.

[24] Independent Advisory Group on a Data Revolution for Susteinable Development. *A World that Counts: Mobilising the Data Revolution for Sustainable Development.* Naciones Unidas Comisión Económica para América Latina y el Caribe (CEPAL), 2014.

[25] S. Román, P. J. Cuestas, and P. Fenollar. An examination of the interrelationships between selfesteem, others' expectations, family support, learning approaches and academic

achievement. *Studies in Higher Education*, 33(2):127–138, 2008.

[26] Henna Rytkönen, A. Parpala, S. Lindblom-Ylänne, V. Virtanen, and L. Postareff. Factors affecting bioscience students' academic achievement. *Instructional Science*, 40(2):241–256, 2012.

[27] N. Sael, A. Marzak, and H. Behja. Multilevel clustering and association rule mining for learnersṕrofiles analysis. *International Journal of Computer Science Issues*, 10(1):188–194, 2013.

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

[29] F. Tobar, L. Yacher, R. Paredes, and M. E. Orchard. Anomaly detection in power generation plants using similarity-based modeling and multivariate analysis. In *Proceedings of the 2011 American Control Conference*, pages 1940–1945. IEEE, 2011.

[30] D. Wilson, D. Jones, M. J. Kim, C. Allendoerfer, R. Bates, J. Crawford, T. Floyd-Smith, M. Plett, and N. Veilleux. The link between cocurricular activities and academic engagement in engineering education. volume 104, pages 625–651, 2014.

Apéndice A

Estadística descriptiva

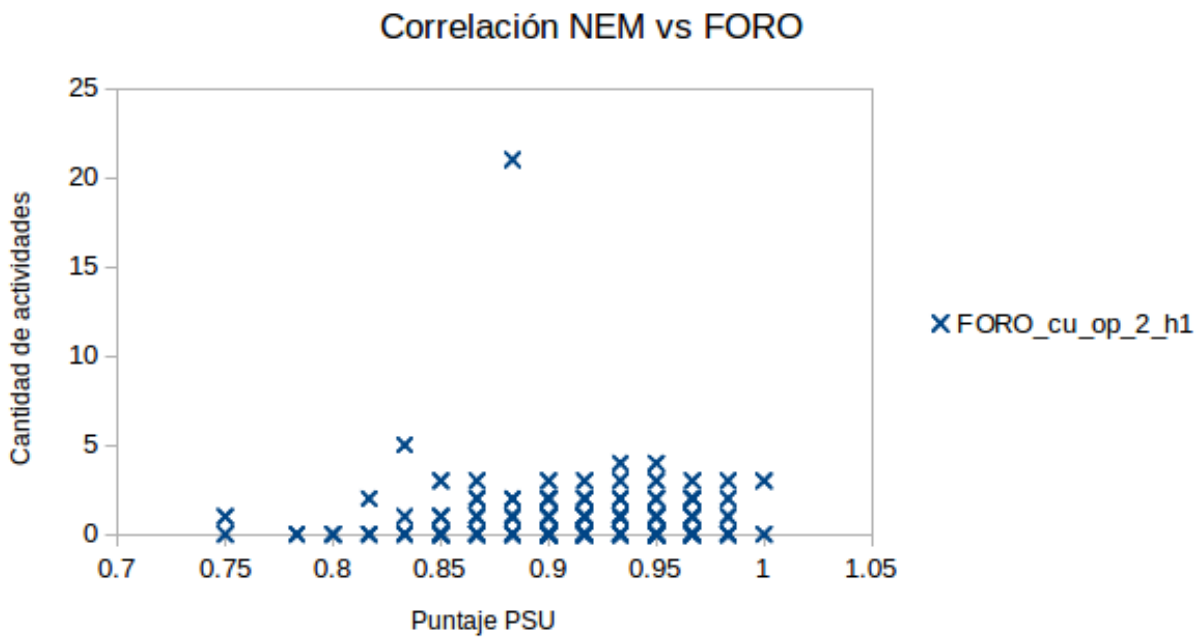

Figura A.1: *Correlación entre las notas de enseñanza media y el número de publicaciones en el foro antes del control 1. Cada marca representa a un estudiante.*

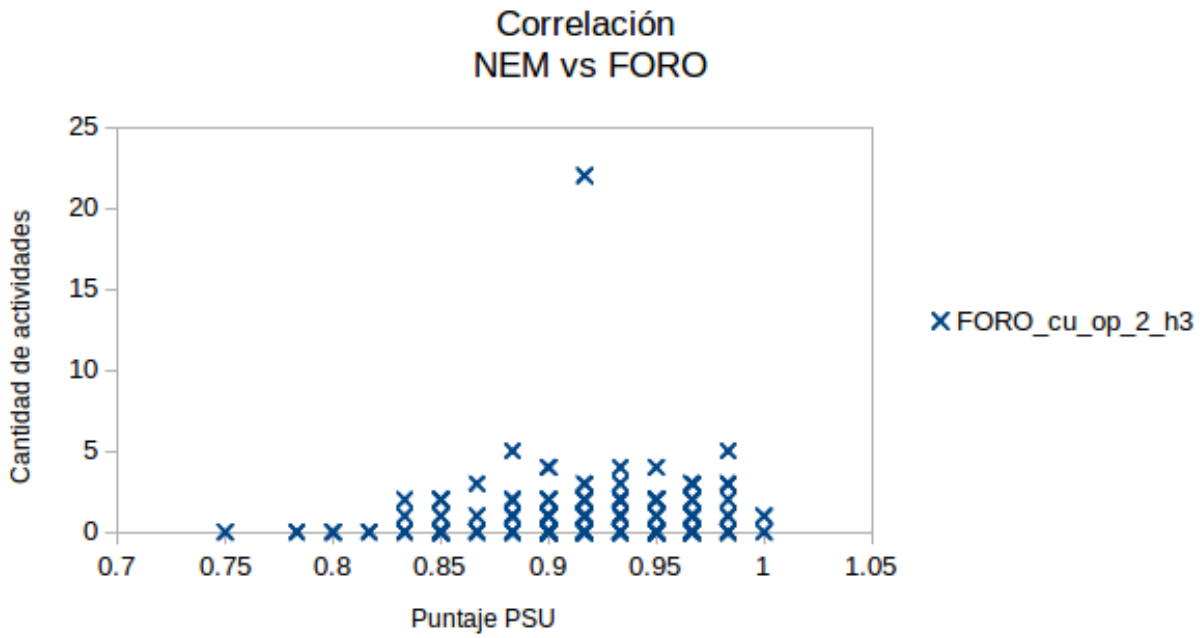

Figura A.2: *Correlación entre las notas de enseñanza media y el número de publicaciones en el foro entre el control 2 y el control 3. Cada marca representa a un estudiante.*

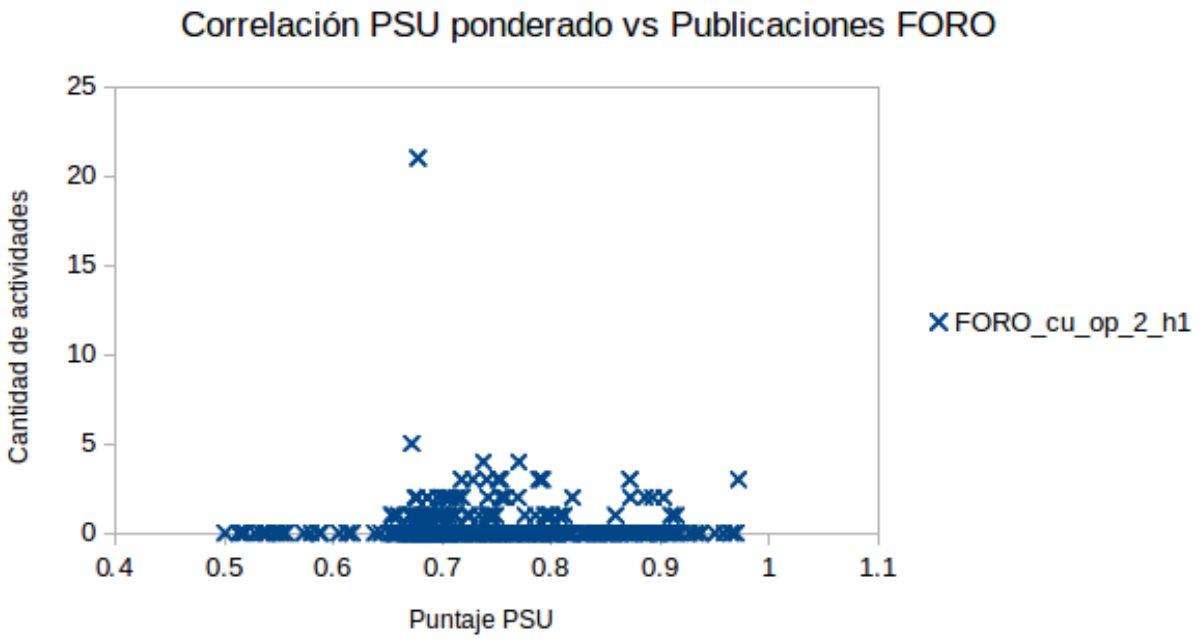

Figura A.3: *Correlación entre el puntaje PSU ponderado y el número de publicaciones en el foro antes del control 1. Cada marca representa a un estudiante.*

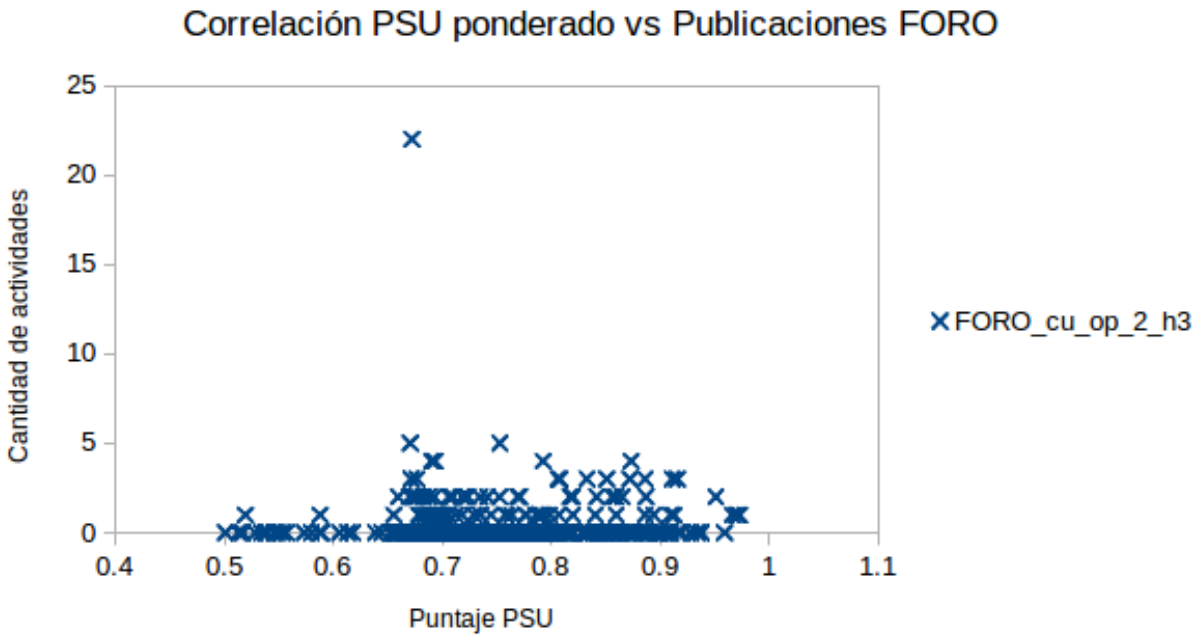

Figura A.4: *Correlación entre el puntaje PSU ponderado y el número de publicaciones en el foro entre el control 2 y el control 3. Cada marca representa a un estudiante.*