

Tabla de Contenido

1. Introducción	1
1.1. Objetivos	4
1.2. Alcances	5
1.3. Estructura General	5
2. Revisión Bibliográfica	6
2.1. Learning Analytics	6
2.2. Teorías del Aprendizaje	8
2.3. Herramientas de Análisis de Datos	9
2.3.1. Sistemas de reconocimiento de patrones	9
2.3.2. Selección de variables mediante <i>Partial Least Squares</i>	9
2.3.3. Reducción de dimensión mediante <i>Principal Component Analysis</i>	11
2.3.4. <i>Similarity Based Modeling</i>	11
2.3.5. Matriz de confusión, <i>Recall</i> y <i>Precision</i>	12
3. Metodología	14
4. Implementación de Modelos Basados en Similitud	19
4.1. Base de Datos	19
4.2. Pre-procesamiento	21
4.2.1. <i>Logs</i> de U-cursos	21
4.2.2. Acotación de variables	22
4.3. Selección de variables mediante <i>PLS</i> y <i>PCA</i>	24
4.3.1. Selección de variables con <i>Partial Least Squares</i>	24
4.3.2. <i>Principal Component Analysis (PCA)</i>	24
4.4. Implementación de <i>SBM</i>	25
5. Resultados	27
5.1. Pre-procesamiento	27
5.1.1. Selección de variables con <i>PLS</i>	27
5.1.2. Variables <i>PCA</i>	28
5.2. Resultados de la clasificación con <i>SBM</i>	29
6. Discusión	35
7. Conclusión	38

Bibliografía	40
Apéndices	43
A. Estadística descriptiva	43

Índice de Tablas

2.1. Matriz de confusión	13
4.1. Descripción de las variables correspondientes a los antecedentes de las cohortes 2013 y 2014.	20
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.	20
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.	21
5.1. Variables eliminadas durante <i>PLS</i> . 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. .	32
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.	33
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.	34

Índice de Ilustraciones

3.1.	Esquema de la metodología.	15
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 <i>online</i> acumulada por un estudiante particular. La línea 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.	17
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.	22
4.2.	Histograma de las variables relacionadas con los <i>logs</i> 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.	23
4.3.	Histograma de las distancias entre observaciones (estudiantes). La línea verde muestra el parámetro σ	26
5.1.	Pesos de variables en los primeros 2 vectores de carga.	28

5.2.	Comparación de clasificadores en términos de <i>recall</i> 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 <i>online</i> 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 <i>recall</i> 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 <i>logs</i> de U-cursos. En cambio, la línea naranja muestra el desempeño del clasificador al no incluir la información de la plataforma <i>LMS</i> , manteniendo las notas de controles y antecedentes socio-demográficos y de admisión.	30
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 <i>online</i> 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 <i>logs</i> de U-cursos. En cambio, la línea naranja muestra el desempeño del clasificador al no incluir la información de la plataforma <i>LMS</i> , manteniendo las notas de controles y antecedentes socio-demográficos.	31
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.	43
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.	44
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.	44
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.	45