PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

PROPUESTA METODOLÓGICA DE MINERÍA DE PROCESOS PARA ANÁLISIS DE DATOS EDUCACIONALES EN MOOCS

NICOLÁS IVÁN MORALES MACAYA

Tesis para optar al grado de
Magíster en Ciencias de Ingeniería

Profesor Supervisor:
MAR PÉREZ SANAGUSTÍN

Santiago de Chile, Julio 2018

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

PROPUESTA METODOLÓGICA DE MINERÍA DE PROCESOS PARA ANÁLISIS DE DATOS EDUCACIONALES EN MOOCS

NICOLÁS IVÁN MORALES MACAYA

Tesis presentada a la Comisión integrada por los profesores:
MAR PÉREZ SANAGUSTÍN
JORGE MUÑOZ GAMA
ÁNGEL ABUSLEME HOFFMAN
SERGIO CELIS GUZMÁN
IGNACIO VARGAS CUCURELLA

Para completar las exigencias del grado de
Magíster en Ciencias de Ingeniería

Santiago de Chile, Julio 2018

A Cecilia, mi madre, María José, mi hermana,
y mis amigos que me apoyaron cuando lo necesité.

AGRADECIMIENTOS

Los trabajos de tesis son procesos de largo aliento, que requieren mucho esfuerzo y dedicación, más la voluntad para superar las incertezas propias del trabajo investigativo, y de la vida misma. Tareas de este calibre requieren el apoyo de múltiples personas, y en mi caso, sé que no hubiera posible sin el apoyo recibido de quienes me rodean, ofrezco mi más sincero agradecimiento a todos quienes me han brindado el apoyo necesario a lo largo de este periodo en que he sido estudiante del programa de magíster.

Agradezco a Mar, mi profesora guía, por ser alguien que me ha enseñado mucho, de quien sigo aprendiendo, por haber estado presente y acompañarme durante este proceso, y por haberme dado la oportunidad de trabajar bajo su supervisión.

A mis compañeros del equipo de investigación, el T4D Lab, en especial a Jorge, que ha sido una ayuda invaluable en este trabajo y con quien he podido contar siempre, también a Ronald y a los demás chicos, de quienes he aprendido mucho.

A mi familia, en especial a mi madre, quien a pesar de no entender bien lo que hacía o algunas de mis decisiones, nunca dejó de darme alientos para terminar este trabajo.

A mis compañeros y amigos de la universidad, y de la vida, a Rodrigo Dedes, a Patricia, a Andrés, a los chicos del capítulo estudiantil del IEEE, Rodrigo Henríquez, Sebastián, Simón, Gabriel, Renzo, Marie, Kevin, Diego y todos los demás que han pasado por esa oficina, también al Sebastián Godoy, a Emilio, mi estancia en la universidad no hubiera sido lo mismo sin ellos, todos han sido excelentes compañeros y amigos, y sé que siempre he podido contar con ustedes, probablemente se me pase más de alguien escribiendo esto y pido disculpas de antemano por ello, pero es mucha la gente que he conocido en mi paso por la universidad, y es de lo más valioso que me he encontrado en mi paso por esta. También doy las gracias a los compañeros de diversos proyectos e iniciativas en las que participé, como el GEAC, Kunlabora y Crecer, que le dieron sentido a mi paso por la

universidad, estoy convencido de que uno no puede asistir solo a estudiar en esta, contar con la posibilidad de estudiar en la universidad y realizar un postgrado es un privilegio, y siento que es un deber trabajar más allá del proyecto individual y personal con los conocimientos adquiridos.

ÍNDICE GENERAL

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

Los MOOC (cursos masivos abiertos en línea han resultado ser una de las principales innovaciones en educación superior de la última década, permitiendo a instituciones educativas distribuir contenidos de calidad a miles de estudiantes alrededor del mundo. Las plataformas registran en detalle la actividad de los estudiantes, abriendo nuevas posibilidades de investigar la actividad y rendimiento de estos, profundizando así la investigación realizada en *Learning Analytics* para explicar cómo desarrollan su proceso de aprendizaje los estudiantes. Dentro de las distintas técnicas de análisis existentes, la Minería de Procesos (PM) destaca al facilitar el descubrimiento de modelos de procesos de aprendizaje que representan la secuencia de las interacciones de los alumnos con los materiales del curso, lo que permite incluir una visión global del proceso de aprendizaje en el análisis. El objetivo de este trabajo de tesis es desarrollar una propuesta metodológica basada en técnicas de minería de procesos y la adaptación de metodologías existentes, para la extracción y análisis del comportamiento de los estudiantes de MOOCs, además de su validación en un caso de estudio para extraer comportamiento de los estudiante, donde se analizaron datos de 3458 participantes de 3 cursos de la plataforma Coursera, extrayendo seis patrones de interacción que dan cuenta de distintos tipos de sesiones de trabajo que los estudiantes realizan en el curso y clasificando a los estudiantes en tres grupos en base al tipo de actividad que han desarrollado.

Palabras claves: Cursos Masivos Abiertos en Línea, Analíticas de Aprendizaje, Minería de Procesos, Proceso de Aprendizaje, Entornos de Aprendizaje en Línea, Datos Educacionales, Minería de Datos Educacional, Minería de Procesos Educacional.

ABSTRACT

MOOCs (mass open online courses) have proven to be one of the major innovations in higher education in the last decade, enabling educational institutions to distribute quality content to thousands of students around the world. The platforms record in detail the activity of students, opening new possibilities to investigate their activity and performance, thus deepening the research carried out in *Learning Analytics* to explain how students develop their learning process. Within the different existing analysis techniques, Process Mining (PM) stands out by facilitating the discovery of models of learning processes that represent the sequence of students' interactions with course materials, allowing a global vision of the learning process to be included in the analysis. The objective of this thesis work is to develop a methodological proposal based on process mining techniques and the adaptation of existing methodologies, for the extraction and analysis of the behaviour of MOOC students, in addition to its validation in a case study to extract student behaviour, where data from 3458 participants of 3 courses of the Coursera platform were analysed, extracting six interaction patterns that account for different types of work sessions that students carry out in the course and classifying students into three groups based on the type of activity they have developed.

Keywords: Massive Open Online Courses, *Learning Analytics*, Process Mining, Learning Processes, Online Learning Environments, Educational Data, Educational Data Mining, Educational Process Mining.

1. INTRODUCCIÓN

1.1. Motivación

Los MOOC (cursos masivos y abiertos en línea, de su sigla en inglés) han resultado ser una de las principales innovaciones en educación superior de la última década, permitiendo a las instituciones educativas distribuir contenidos de calidad a miles de estudiantes alrededor del mundo (Breslow et al., 2013; Daradoumis et al., 2013), el concepto MOOC es utilizado por primera vez en 2008, por Cormier (2008). Dentro de las plataformas pioneras de MOOC, tenemos a Coursera, edX, Udacity, FutureLearn y MiríadaX, entre otras, las que empezaron con un modelo consistente en cursos gratuitos, basados en vídeos, los que siguen una estructura con fechas de inicio y finalización para cada curso. Los MOOC disponibles inicialmente se basaban en la web, sin embargo, las iniciativas MOOC han invertido en nuevos desarrollos para apoyar una experiencia de aprendizaje móvil e ininterrumpida (Wong y Looi, 2011; de Waard et al., 2011; Wong, 2013; Milrad et al., 2013; Sharples et al., 2015; de Waard et al., 2016).

Dentro de los desarrollos más destacables contamos que Coursera transformó su primera plataforma en una plataforma de MOOC bajo demanda, modelo que ha sido adoptado también por otras iniciativas. Los cursos bajo demanda permiten a los estudiantes tomar un MOOC siempre que lo deseen, sin tener que empezar y terminar en una fecha determinada (*Coursera Update: Striking a Balance with Start Dates and Deadlines | Coursera Blog*, 2013). Por otra parte, en el año 2013 fue cuando edX lanzó su plataforma de código abierto, Open edX. Esta plataforma ofrece a todas las universidades la oportunidad de publicar por su cuenta sus propios MOOC y aprovechar los recursos de estos dentro de sus iniciativas en el campus, tanto para la investigación como para la docencia, y tener control sobre de sus propios datos (*Stanford online coursework to be available on new open-source platform*, 2013).

La implementación y evolución de las plataformas MOOC ha abierto un nuevo abanico de escenarios de aprendizaje basados en el uso de MOOCs. Estos escenarios se basan en

distintas metodologías, desde las puramente online, a las mixtas (o híbridas) como por ejemplo el *flipped classroom* (Y. Zhang, 2013; Kloos et al., 2015; Ho et al., 2015; Soffer y Cohen, 2015; Pérez-Sanagustín et al., 2017).

Este tipo de escenarios educativos son cada vez más frecuentes y masivos en las universidades a nivel mundial, lo que ha conllevado, en los últimos años, y desde que aparecen los cursos MOOCs en el 2008, la recolección de grandes datos masivos para la educación. Las plataformas de aprendizaje digital recopilan registros detallados del comportamiento, el rendimiento y otros tipos de interacción de cada alumno. En particular, los MOOC, dado su carácter masivo con cursos de miles de estudiantes, son una fuente importante de datos sobre el comportamiento de los alumnos y permiten a la investigación comprender mejor cómo aprenden las personas en entornos de aprendizaje en línea (Breslow et al., 2013; Cooper y Sahami, 2013; Daradoumis et al., 2013).

Este incremento masivo de los datos y registros de datos de granularidad fina ha dado pie a una nueva área de investigación en el área de tecnología educativa denominada *Learning Analytics* (Dietze, Siemens, Taibi y Drachsler, 2016). El *Learning Analytics* puede definirse como la recolección, análisis y aplicación de los datos acumulados para evaluar el comportamiento de las comunidades educativas. Ya sea mediante el uso de técnicas estadísticas y modelos predictivos, visualizaciones interactivas o taxonomías y marcos de trabajo, el objetivo principal que persigue es optimizar el rendimiento tanto de los estudiantes como del profesorado, para refinar las estrategias pedagógicas, racionalizar los costes institucionales, para determinar el compromiso de los estudiantes con el curso material, para resaltar a los estudiantes con dificultades potenciales (y para alterar la pedagogía en consecuencia) afinar los sistemas de calificación mediante análisis en tiempo real y permitir que los instructores juzguen su propia eficacia educativa. También otros objetivos posibles pueden ser el producir retroalimentación y recomendaciones a los distintos actores educativos, o entregar pronósticos que permitan mejorar la toma de decisiones de estos. En todos los casos, el *Learning Analytics* ofrece a las partes interesadas una visión de lo que está ocurriendo desde el día 1 hasta el día X de un curso dado, independientemente del

tipo de actividad que se esté llevando a cabo. En resumen, la analítica de aprendizaje se define en términos generales como el esfuerzo por mejorar la enseñanza y el aprendizaje a través del análisis específico de los datos demográficos, de rendimiento y comportamiento de los estudiantes (Elias, 2011; Fritz, 2011; Larusson y White, 2014).

Enmarcados en esta área, muchos autores se han dedicado a proponer métricas y análisis para entender cómo extraer información sobre el comportamiento de los estudiantes en los MOOCs, y desde antes del desarrollo de este concepto, en los LMS (sistema de gestión de aprendizaje, de su sigla en inglés). Algunos ejemplos relevantes para el contexto MOOC son: Hadwin et al. (2007), quien estudió las secuencias de las acciones de SRL (auto-regulación del aprendizaje, de su sigla en inglés) mediante gráficos de transiciones de las actividades desarrolladas, utilizando una herramienta de software en línea denominada gStudy, donde se determinó posibles perfiles de SRL basado en la actividad sobre el software desarrollado; Bogarín et al. (2014), quienes en su propuesta utilizaron técnicas de clusterización para agrupar conjuntos de secuencias realizadas por 84 estudiantes de pre-grado que utilizaron la plataforma Moodle 2.0 como parte de un curso en línea, y compararon los modelos de procesos obtenidos; Beheshitha et al. (2015), quienes en su propuesta utilizaron un algoritmo difuso para examinar la relación que existente entre las aptitudes reportadas en un cuestionario y las estrategias de SRL cognitivas desplegadas con un grupo de 22 estudiantes utilizando el software nStudy, encontrando como resultado en base a la actividad en el software dos grupos de estudiantes (superficiales y profundos); y también Mukala, Buijs, Leemans, y van der Aalst (2015), quienes exploraron utilizando técnicas de Minería de Procesos (PM) en un MOOC con 43,218 estudiantes inscritos, el comportamiento de estos estudiantes durante el proceso de aprendizaje. Sin embargo, un número masivo de datos no significa necesariamente que sea sencillo para validar o construir sobre la base de las teorías educativas existentes. En particular, y como señalan J. Lodge y Lewis (2012) el acceso a información crítica sobre el comportamiento y los procesos de aprendizaje de los alumnos suele ser limitado. Los métodos basados en datos pueden extraer rápidamente patrones de actividades de alumnos a lo largo de un curso,

pero sigue siendo un reto el interpretar y entender cómo estos patrones se relacionan con la teoría.

Un enfoque para mejorar la interpretación de estos datos y relacionarlos con la teoría es triangular con otras fuentes (es decir, adoptar un enfoque de métodos mixtos). Por ejemplo, los datos de actividades registradas en MOOCs, que capturan las interacciones reales de los alumnos, pueden combinarse con datos de instrumentos auto-reportados como cuestionarios o sesiones de pensamiento en voz alta (Eynon, 2013; Bannert et al., 2014) , o datos de fuentes externas como tracking ocular (Trevors et al., 2016).

Estos estudios son ejemplos de un avance hacia el uso de metodologías más complejas para la interpretación de los datos masivos extraídos en MOOCs. Sin embargo, aún sigue siendo un reto el explorar formas de conectar la teoría educativa con los métodos basados en datos con los datos de auto-informe y de comportamiento para comprender mejor cómo se comportan y aprenden los alumnos en entornos digitales (J. M. Lodge y Corrin, 2017).

En este contexto, las técnicas de minería de procesos, de ahora en adelante PM (sigla en inglés de Process Mining) presentan una oportunidad para enfocar el análisis de datos masivos educativos en cómo es el comportamiento de los estudiantes de forma directa desde los datos, lo que puede ser contrastado con las teorías educativas existentes y modelos formulados por los investigadores. En ese sentido, las técnicas de PM facilitan el descubrimiento de modelos de procesos de aprendizaje que representan la secuencia de las interacciones de los alumnos con los materiales del curso (Van der Aalst, 2011), y también proporcionan una metodología probada para extraer, analizar y visualizar las trazas de la interacción de los alumnos (Jivet, 2016; Mukala, Buijs, y Van Der Aalst, 2015; Romero et al., 2016). Dentro de las diversas técnicas de análisis posibles para estos datos, el enfoque de PM proporciona un puente entre la minería de datos y el modelado y análisis de procesos, donde el PM como una sub-disciplina de la minería de datos añade la visión orientada al proceso del procedimiento de minería de datos (Van der Aalst et al., 2004a).

En los últimos años, la minería de datos ha sido ampliamente aplicada con éxito para encontrar patrones interesantes de datos recopilados en entornos educativos (Romero y Ventura, 2017), sin embargo, las técnicas de EDM (minería de datos educacional, de su sigla en inglés) se centran en dependencias de datos o patrones simples y no proporcionan una representación visual de todo el proceso de aprendizaje, es decir, no se centra en el proceso en su conjunto, con lo cual, considerando las potencialidades recién presentadas en relación a las técnicas de PM, surge como área de investigación el EPM (minería de procesos educacional, de su sigla en inglés), el cual es considerando como un área dentro del EDM, con el fin de explicitar conocimiento no expresado directamente en los datos y facilitar una mejor comprensión del proceso educativo (Bogarín et al., 2018). En particular para cursos MOOC, el uso de PM destaca al proporcionar indicaciones útiles en términos de información y orientación a considerar para realizar intervenciones a fin de mejorar la calidad y la entrega de este tipo de cursos (Mukala, Buijs, Leemans, y van der Aalst, 2015; Sonnenberg y Bannert, 2015).

Para avanzar en el área de *Learning Analytics* aplicado a datos MOOC y, más concretamente, para avanzar en el área del EPM, este trabajo propone explorar cómo entender mejor, y caracterizar el proceso de aprendizaje en un contexto MOOC aprovechando el potencial de las técnicas de PM. Concretamente, se propone adaptar metodologías existentes de PM, para luego aplicar esta propuesta metodológica a un caso en particular con datos de cursos de la plataforma Coursera, utilizando técnicas de descubrimiento de procesos, para así posteriormente analizar la actividad y a los estudiantes del curso en base al modelo de proceso obtenido.

1.2. Objetivos

El objetivo general de esta tesis es desarrollar una propuesta metodológica basada en técnicas de minería de procesos para la extracción y análisis del comportamiento de los estudiantes de MOOCs.

Los objetivos específicos corresponden a:

1. Adaptación de metodologías existentes de PM para modelar y analizar la actividad de estudiantes en un contexto MOOC.
2. Aplicar la propuesta metodológica a un caso real para extraer comportamiento de los estudiantes en un MOOC, para así validar esta propuesta.

1.3. Preguntas de investigación

Este trabajo tiene como objetivo explorar cómo las técnicas de PM contribuyen a analizar los procesos educativos de datos masivos. Concretamente, se plantean las siguientes preguntas de investigación:

- RQ1. ¿Es posible adaptar y proponer una metodología basada en técnicas de PM para estudiar el comportamiento de los estudiantes a partir de datos de una plataforma MOOC?
- RQ2. ¿Qué variaciones debo realizar en la metodología de PM para extraer patrones de comportamiento y perfiles de estudiante a partir de datos de varios MOOCs?

1.4. Metodología

El trabajo desarrollado en esta tesis, primero pasa por una búsqueda bibliográfica relativa a las técnicas y metodologías existentes en PM, además de sus aplicaciones en el ámbito de la educación. Una vez clasificadas y depuradas las fuentes, se procede a seleccionar la metodología PM^2 de minería de procesos como base para la propuesta metodológica para abordar investigaciones sobre procesos no estructurados, como es el caso del contexto MOOC.

Una vez seleccionada esta metodología como base, se procede a determinar qué fases y actividades de esta es necesario adaptar, y para qué actividades es necesario presentar definiciones específicas del contexto MOOC, de forma que se pueda completar la propuesta metodológica.

Con la propuesta metodológica ya constituida, ésta es aplicada a un caso de estudio con datos de la plataforma Coursera, para comprobar la aplicabilidad de la metodología así como su efectividad, por medio de realizar un análisis exhaustivo del comportamiento de los estudiantes en los cursos estudiados.

1.5. Organización de la tesis

La tesis está organizada de la siguiente manera.

El **capítulo 2**, "Minería de Procesos (PM) aplicada a la educación", presenta el área del *Educational Process Mining* (EPM) como una aplicación de PM a los datos educacionales, una recopilación de los principales trabajos de EPM relacionados con el contexto MOOC y plataformas similares, como los *Learning Management Systems* (LMSs), y algunas áreas de estudio dentro de *Educational Data Mining* (EDM) que están relacionadas con el EPM, para luego explicar los conceptos más importantes de PM, las técnicas y herramientas disponibles para aplicar PM en el estudio de procesos de aprendizaje, y finalmente las problemáticas más comunes al momento de implementar un proyecto de PM.

El **capítulo 3**, "Propuesta metodológica de Minería de Procesos para análisis del proceso de aprendizaje en cursos MOOC", presenta primero la metodología PM^2, sus características generales, una descripción por etapas de esta y cómo estas etapas de la metodología serán adaptadas al contexto MOOC.

El **capítulo 4**, "Caso de estudio: Aplicación de la metodología para el estudio del proceso de aprendizaje en cursos MOOC de la plataforma Coursera", presenta una aplicación de la propuesta metodológica, comenzando con una descripción del set de datos y cursos

considerados para la investigación, la presentación de un instrumento que fue aplicado a los estudiantes para auto-reporte de su nivel de auto-regulación del aprendizaje, y luego las etapas y definiciones requeridas para el descubrimiento del modelo de proceso. Posteriormente, se presenta el análisis desarrollado sobre los datos en base al modelo de proceso generado, tanto en la caracterización de los patrones de actividad encontrados, como los tipos de estudiantes que es posible clasificar en base a la actividad, además de la relación de esta actividad con indicadores de auto-regulación del aprendizaje y la aprobación en el curso.

El **capítulo 5**, "Conclusiones y trabajo futuro", presenta algunas conclusiones y líneas de trabajo futuras relacionadas a lo presentado en esta tesis. Se incluyen algunos comentarios finales y opiniones personales.

1.6. Contribuciones

Esta tesis tiene dos contribuciones principales: (1) propuesta metodológica de 4 fases basada en la metodología PM^2 de Minería de Procesos, con enfoque a la aplicación en plataformas MOOC; y (2) un caso de estudio ejemplar que utiliza datos de la plataforma Coursera, en donde se aplica la propuesta metodológica para el análisis de los patrones de actividad observados y perfiles de estudiantes en base a la actividad registrada en los cursos.

2. MINERÍA DE PROCESOS (PM) APLICADA A LA EDUCACIÓN

En este capítulo se presentan los principales aspectos de PM, y las particularidades del área de investigación conocida como *Educational Process Mining* (EPM), que corresponde a la aplicación de PM en datos educacionales, además de mencionar las áreas relacionadas con el EPM, las técnicas y herramientas disponibles para su aplicación, y para finalizar, las principales problemáticas en la aplicación de EPM en general y a un contexto MOOC.

2.1. Educational Process Mining (EPM)

Una definición consensuada por la comunidad, sobre qué es la minería de procesos, es la que se puede encontrar en el *Process Mining Manifesto* (Van Der Aalst et al., 2011):

> La minería de procesos es una disciplina de investigación relativamente joven que se ubica entre la inteligencia computacional y la minería de datos, por una parte, y la modelación y análisis de procesos, por otra. La idea de la minería de procesos es descubrir, monitorear y mejorar los procesos reales (i.e., no los procesos supuestos) a traves de la extraccion de conocimiento de los registros de eventos ampliamente disponibles en los actuales sistemas (de información). (pp. 1-2)

La mayor parte del trabajo realizado en PM se ha concentrado en los sistemas de flujo de trabajo (de negocios) y en el descubrimiento de las representaciones de flujos de trabajo en redes de Petri (Trcka y Pechenizkiy, 2009), donde al PM también se le ha denominado *Workflow Mining* (WM). La denominación del área como *Workflow Mining* era más frecuente en los primeros trabajos del área, como por ejemplo en el artículo que presenta el algoritmo Alpha de descubrimiento de procesos (Van der Aalst et al., 2004b), el cual es la base para diversos algoritmos de descubrimiento utilizados actualmente, algunos de los cuales serán mencionados más adelante en este capítulo.

Sin embargo, en los últimos años el enfoque de PM ha sido aplicado a otros dominios distintos a los procesos de negocios, como por ejemplo la salud (Rojas et al., 2016) y la educación (Jivet, 2016; Mukala, Buijs, y Van Der Aalst, 2015; Romero et al., 2016). En particular, la aplicación de PM sobre datos educacionales, ha sido denominada como *Educational Process Mining* (EPM), donde lo que caracteriza el EPM, es que este trabaja en base a los registros de eventos generados por los *virtual learning enviroments* (VLEs) (Bogarín et al., 2018).

Los trabajos desarrollados en EPM requieren un marco teórico adecuado, que permita guiar el cómo construir los modelos de proceso, para que éstos efectivamente permitan responder preguntas requeridas por los educadores y diseñadores instruccionales en torno a cómo ocurren y cómo es posible mejorar los procesos de aprendizaje. Un enfoque inductivo, solamente basado en datos, si bien permite extraer de manera relativamente fácil información sobre el proceso, tiene un valor limitado al buscar regularidades y patrones sin un marco conceptual rector (Bogarín et al., 2018). El enfoque de EPM es capaz de construir modelos de procesos educativos completos y compactos que son capaces de reproducir todos los comportamientos observados, comprobar si el comportamiento de modelado coincide con el comportamiento observado, proyectar la información extraída de los registros en el patrón para hacer explícito el conocimiento tácito, y facilitar una mejor comprensión del proceso (Trcka y Pechenizkiy, 2009).

El EPM se ha aplicado en diversos contextos educativos, como lo son los *Learning Management Systems* (LMSs), *Computer-supported collaborative learning* (CSCL), cursos de entrenamiento profesional, evaluaciones basadas en computadora, entre otros (Bogarín et al., 2018). Dentro de estas áreas, las aplicaciones en LMS son las más relevantes para este trabajo, dado que las plataformas MOOC más populares como Coursera y edX, soportan sus cursos dentro de sus propios LMS, con la diferencia de que en este caso los cursos ofrecidos son abiertos y masivos (de cientos o miles de estudiantes).

Dentro de los trabajos relacionados con MOOCs y LMSs, hay múltiples trabajos recopilados en una reciente revisión del estado del arte en EPM, hecha por Bogarín et al.

(2018). Trcka y Pechenizkiy (2009) ejemplificaron la aplicabilidad de PM y analizaron su potencial para extraer información de los LMSs, considerando solamente las trazas de las evaluaciones rendidas por los estudiantes. En Bogarín et al. (2014), los autores utilizaron datos de los registros de la plataforma Moodle y propusieron clusterizar los estudiantes para obtener modelos de proceso más específicos y precisos en relación al comportamiento de los distintos estudiantes. En un entorno similar Reimann et al. (2014) proponen el uso de trazas para estudiar el aprendizaje autorregulado en un ambiente de hipermedios, basados en principios teóricos y técnicas de PM. En base a estos principios teóricos, se detectan diferencias en las frecuencias de eventos de aprendizaje autorregulado utilizando técnicas de PM, y se encontró que los estudiantes exitosos demuestran más eventos de aprendizaje y regulación (Bannert et al., 2014). Otro grupo de investigación, utilizó técnicas de PM para para rastrear y analizar los hábitos de aprendizaje de los estudiantes en base a en los datos del MOOC, donde se encontró que los estudiantes exitosos en completar el curso siguen un patrón de observación estructurado secuencialmente mientras que los estudiantes que no logran completar el curso son impredecibles y tienen procesos mal estructurados (Mukala, Buijs, Leemans, y van der Aalst, 2015). Posteriormente este grupo por medio de técnicas de análisis de conformidad sobre los datos de estos MOOC, extrajeron patrones de aprendizaje de este (Mukala, Buijs, y Van Der Aalst, 2015). También Emond y Buffett (2015) aplicaron técnicas de descubrimiento de procesos y clasificación de secuencias para modelar y dar soporte al aprendizaje autorregulado en en entornos heterogéneos de contenido de aprendizaje, actividades, y redes sociales, esto utilizando un conjunto de datos de actividades de aprendizaje semiestructuradas, tomadas del *DataShop* en el Pittsburgh Science of Learning Centre. Por último, Vidal et al. (2016) usaron registros de un VLE para extraer la estructura del flujo de aprendizaje utilizando PM, y obtener las reglas subyacentes que controlan la capacidad de adaptación del aprendizaje de los estudiantes mediante árboles de decisión.

Otra investigación cuya publicación es más reciente, por lo que no fue incluida en la revisión del estado del arte mencionada, es la realizada por Maldonado-Mahauad et al. (2018), donde se aplican técnicas de PM para obtener modelos de proceso a un nivel de

sesión de trabajo de los estudiantes en cursos MOOC de la plataforma Coursera. Además, estos procesos se clasifican para obtener distintos perfiles de estudiante y analizar cómo estos perfiles se relacionan con el éxito en el curso y los procesos de aprendizaje autorregulado que desarrollan. El autor de este trabajo participó en el artículo previamente mencionado y parte de este artículo se presenta como parte del caso de estudio de la propuesta metodológica de PM para el análisis de datos educacionales en cursos MOOC.

2.2. Áreas relacionadas al EPM

Dentro de las distintas áreas de la minería de datos, existen algunas que son cercanas al PM y de particular interés para los contextos educacionales, éstas son: *intention mining* (IM), *sequential pattern mining* (SPM), y *graph mining* (GM) (Bogarín et al., 2018):

- ***Intention Mining* (IM):** IM tiene como objetivo extraer secuencias de las actividades de los usuarios a partir de conjuntos de registros de eventos, para inferir las intenciones de los usuarios relacionados. Un conjunto de actividades corresponde al logro de una intención. Utiliza registros de eventos como entrada y produce modelos de procesos intencionales (Khodabandelou et al., 2013).[1]
- ***Sequential pattern mining* (SPM):** SPM es una técnica muy utilizada en el entorno del DM para descubrir sub-secuencias comunes. El SPM tiene como objetivo encontrar las relaciones entre las ocurrencias de eventos secuenciales, es decir, encontrar si existe un orden específico de ocurrencias (Agrawal y Srikant, 1995; Nesbit et al., 2007). Dentro de las técnicas relacionadas al SPM se pueden encontrar el *lag sequential analysis* (LAS), *t-pattern analysis* y los modelos de Markov (Bogarín et al., 2018). Todas estas técnicas se adaptan mejor a secuencias recurrentes relativamente cortas y al análisis de transiciones de eventos (Reimann et al., 2009). Si bien las técnicas de SPM se han aplicado ampliamente para analizar los comportamientos de aprendizaje de los estudiantes, son más indicadas cuando

[1]Si bien, no hay investigación de IM aplicada a educación, esta técnica tiene potencial para ser aplicada en este tipo de contextos, dado que es adecuada para inferir las formas de pensar y trabajar de los usuarios, ya que capta el razonamiento humano que hay detrás de las actividades (Bogarín et al., 2018).

se trata de descubrir patrones de comportamiento en serie o más simples que en un proceso, como lo son rutas de actividad particulares de los estudiantes en un curso. Por lo tanto, SPM no es apropiado para descubrir comportamientos de aprendizaje que describan el proceso de aprendizaje en general (Bannert et al., 2014).

- ***Graph mining* (GM):** el objetivo de GM es encontrar todos los sub-grafos frecuentes en un grafo grande o en una base de datos de grafos, está fuertemente relacionado con la minería de datos multi-relacional. Sin embargo, GM busca suministra nuevos principios y algoritmos eficientes para minar las subestructuras topológicas embebidas en los datos gráficos, mientras que por otro lado la minería de datos multi-relacional busca proporcionar principios para el minado y/o aprendizaje de patrones relacionales representados por lenguajes lógicos expresivos. La primera tiene una orientación más bien geométrica y la segunda más lógica y orientada a las relaciones (Washio y Motoda, 2003), el *social network analysis* (SNA) es una de las posibles aplicaciones de GM, por lo que es posible aplicar GM en contextos educacionales para analizar la información de foros de discusión en plataformas educacionales (Bogarín et al., 2018).

2.3. Conceptos principales de PM y adaptación a EPM

El marco de trabajo de PM contiene 4 componentes principales (Van Der Aalst et al., 2011):

- **Mundo:** corresponde a todo lo que interactúa en el mundo real, entiéndase por personas, máquinas, organizaciones, procesos y otros.
- **Sistemas de software:** esto se refiere a todos los programas computacionales que interactúan con las acciones desarrolladas en el mundo real, y que guardan información relativa a estas, esta información permite construir el registro de eventos.
- **Registros de eventos:** corresponden a registros que contienen de forma secuencial las actividades desarrolladas, con la información correspondiente al caso de esta actividad, es decir la instancia del proceso donde ocurre esa actividad. También

pueden incluir información adicional como el usuario que desarrolla la actividad, la marca de tiempo del evento o recursos utilizados o asociados con la actividad realizada. En el caso de los *Process-Aware Information System* (PAIS), el registro de eventos se provee directamente, pero en otros casos, es necesario construir el registro de eventos a partir de la información existente en las distintas bases de datos asociadas a los sistemas de software involucrados (Van der Aalst, 2011).

- **Modelos de procesos:** corresponden a los modelos que indican las posibles secuencias que pueden seguir las actividades de un proceso, existen diversas notaciones para representar estos, como lo son las redes de Petri, las *Bussinness Process Model Network* (BPMN), grafos dirigidos, redes difusas, entre otros. Estos modelos de proceso pueden existir previamente al ser diseñado el proceso, o pueden ser descubiertos desde los datos.

Con estos componentes se pueden aplicar distintos tipos de minería de procesos, donde los tres principales son (Van Der Aalst et al., 2011):

- **Descubrimiento:** corresponde a producir un modelo de procesos a partir de un registro de eventos, sin utilizar información previa, es la técnica más destacada de PM, y existen múltiples algoritmos para ejecutarla.
- **Verificación de conformidad:** consiste en la comparación de un modelo de proceso con un registro de eventos del mismo proceso, de forma que se pueda corroborar si la realidad, tal como está almacenada en el registro de eventos, es equivalente al modelo y viceversa.
- **Mejoramiento:** busca extender o mejorar un modelo de proceso existente usando la información acerca del proceso real almacenada en algún registro de eventos. Mientras la verificación de conformidad mide el alineamiento entre el modelo y la realidad, este tercer tipo de minería de procesos busca cambiar o extender el modelo a-priori. Por ejemplo, al usar marcas de tiempo en el registro de eventos, se puede extender el modelo para mostrar cuellos de botella, niveles de servicio, tiempos de procesamiento, y frecuencias.

En la figura 2.1, se muestra un ejemplo de un registro de eventos, este incluye identificadores para cada caso y evento, el atributo de actividad como clase de evento, y atributos adicionales como la marca de tiempo y los recursos involucrados. Mientras que en la figura 2.2, se muestra un ejemplo de un modelo de procesos, en este caso corresponde a un modelo de procesos descubierto a partir del registro de eventos de la figura 2.1. Mientras que en la figura 2.3, se muestra un diagrama de cómo se relacionan los componentes y los distintos tipos de minería de proceso.

Case id	Event id	Properties				
		Timestamp	Activity	Resource	Cost	...
1	35654423	30-12-2010:11.02	Register request	Pete	50	...
	35654424	31-12-2010:10.06	Examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	Check ticket	Mike	100	...
	35654426	06-01-2011:11.18	Decide	Sara	200	...
	35654427	07-01-2011:14.24	Reject request	Pete	200	...
2	35654483	30-12-2010:11.32	Register request	Mike	50	...
	35654485	30-12-2010:12.12	Check ticket	Mike	100	...
	35654487	30-12-2010:14.16	Examine casually	Pete	400	...
	35654488	05-01-2011:11.22	Decide	Sara	200	...
	35654489	08-01-2011:12.05	Pay compensation	Ellen	200	...
3	35654521	30-12-2010:14.32	Register request	Pete	50	...
	35654522	30-12-2010:15.06	Examine casually	Mike	400	...
	35654524	30-12-2010:16.34	Check ticket	Ellen	100	...
	35654525	06-01-2011:09.18	Decide	Sara	200	...
	35654526	06-01-2011:12.18	Reinitiate request	Sara	200	...
	35654527	06-01-2011:13.06	Examine thoroughly	Sean	400	...
	35654530	08-01-2011:11.43	Check ticket	Pete	100	...
	35654531	09-01-2011:09.55	Decide	Sara	200	...
	35654533	15-01-2011:10.45	Pay compensation	Ellen	200	...
4	35654641	06-01-2011:15.02	Register request	Pete	50	...
	35654643	07-01-2011:12.06	Check ticket	Mike	100	...
	35654644	08-01-2011:14.43	Examine thoroughly	Sean	400	...
	35654645	09-01-2011:12.02	Decide	Sara	200	...
	35654647	12-01-2011:15.44	Reject request	Ellen	200	...
5	35654711	06-01-2011:09.02	Register request	Ellen	50	...
	35654712	07-01-2011:10.16	Examine casually	Mike	400	...
	35654714	08-01-2011:11.22	Check ticket	Pete	100	...
	35654715	10-01-2011:13.28	Decide	Sara	200	...
	35654716	11-01-2011:16.18	Reinitiate request	Sara	200	...
	35654718	14-01-2011:14.33	Check ticket	Ellen	100	...
	35654719	16-01-2011:15.50	Examine casually	Mike	400	...
	35654720	19-01-2011:11.18	Decide	Sara	200	...
	35654721	20-01-2011:12.48	Reinitiate request	Sara	200	...
	35654722	21-01-2011:09.06	Examine casually	Sue	400	...
	35654724	21-01-2011:11.34	Check ticket	Pete	100	...
	35654725	23-01-2011:13.12	Decide	Sara	200	...
	35654726	24-01-2011:14.56	Reject request	Mike	200	...

Figura 2.1. Ejemplo de un registro de eventos. Extraído de Van der Aalst (2011).

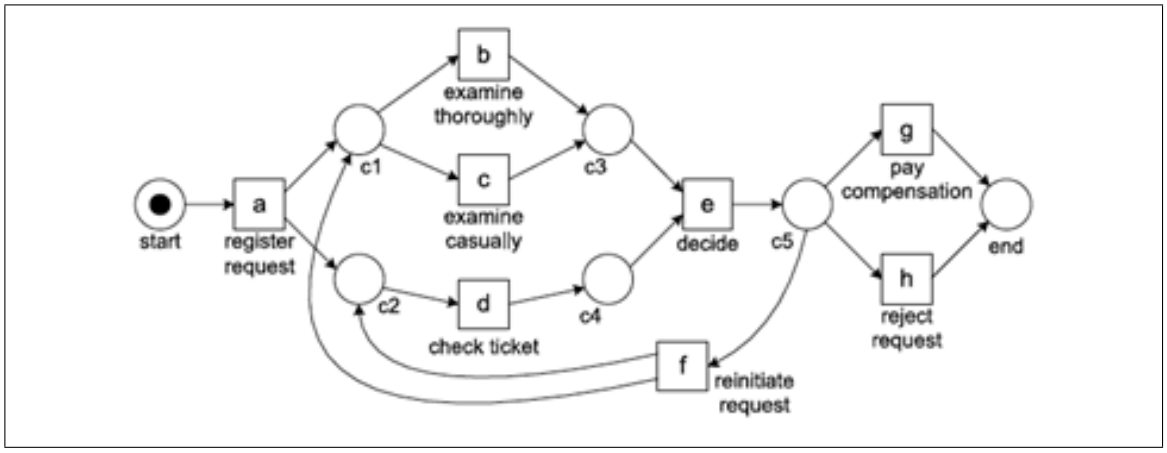


Figura 2.2. Ejemplo de un modelo de procesos. Extraído de Van der Aalst (2011).

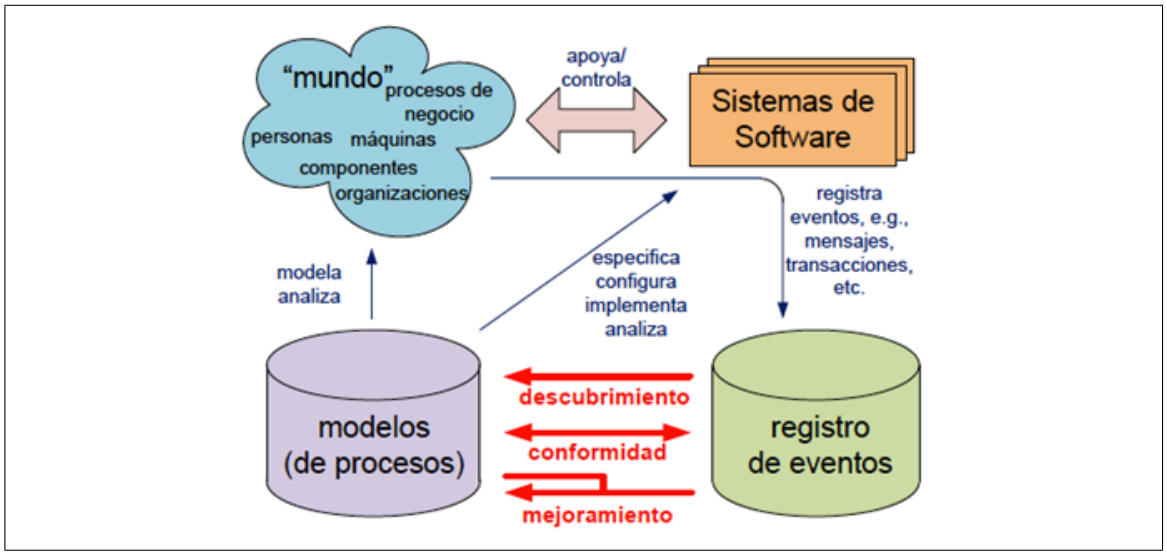


Figura 2.3. Posicionamiento de los tres tipos principales de minería de procesos: (a) descubrimiento, (b) verificación de conformidad, y (c) mejoramiento. Extraído de la versión en español de Van Der Aalst et al. (2011).

Al aplicar PM en el contexto educacional, las principales adaptaciones sobre este marco de trabajo son que el mundo pasa a ser el mundo educacional, y que los sistemas de software son los *virtual learning enviroments* (VLEs) (Cairns et al., 2015; Bogarín et al., 2018).

Dentro del mundo educacional, los principales participantes corresponden a profesores y estudiantes, donde los profesores proveen los recursos utilizados por el estudiante para trabajar y progresar en su aprendizaje, mientras que los estudiantes interactúan con actividades, otros participantes del curso y el mismo sistema. Dentro de esto, los cursos, lecturas, exámenes y otros se utilizan como recursos para los participantes (Bogarín et al., 2018).

Los VLE proveen estructuras y recursos básicos donde ocurren las acciones e interacciones de aprendizaje de los participantes, y registran estos eventos. También proporcionan a profesores e investigadores las herramientas básicas para analizar el aprendizaje de los estudiantes (evolución de las marcas, número de actividades realizadas, participación en el foro, último inicio de sesión, etc.), pero no necesariamente proporcionan herramientas específicas que permitan a los educadores evaluar a fondo el proceso general de aprendizaje del estudiante (Bogarín et al., 2018). En este sentido, usualmente los VLE no pueden ser utilizados como un PAIS donde se pueda obtener directamente el registro de eventos para aplicar técnicas de PM, sino que el registro de eventos debe ser construido de manera ad-hoc a cada VLE y de acuerdo al tipo de proceso que se desea analizar.

Además de los tres tipos principales de PM, existen perspectivas distintas del PM como lo son el control de flujo, la perspectiva organizacional, la de casos y la temporal (Van der Aalst, 2016). En entornos educativos la más utilizada es la perspectiva de control de flujo, que se centra en el ordenamiento de las actividades. El objetivo principal de esta perspectiva es descubrir una descripción ideal de todos los caminos de aprendizaje posibles que se pueden generar cuando los estudiantes navegan a través de un entorno de aprendizaje (Schoonenboom, 2007).

2.4. Técnicas y herramientas disponibles

Dentro de los enfoques de PM más utilizados en educación se encuentran las técnicas de descubrimiento, de verificación de conformidad, el análisis de gráficas de puntos y el análisis de redes sociales (Bogarín et al., 2018).

Dentro de las técnicas de descubrimiento, las más destacadas son:

- **Algoritmo Alpha** (Van der Aalst et al., 2004b), se basa en las transiciones existentes entre una actividad y otra, generando un modelo de procesos que incluye todas las transiciones que es posible encontrar en el registro de eventos sin considerar la frecuencia de estos, solo si existe la transición o no, debido a ello, requiere de registros de eventos ideales sin ruido. Es de los primeros algoritmos que logra descubrir modelos con comportamiento concurrente (Bogarín et al., 2018).
- ***Heuristic Miner*** (A. J. Weijters y Van der Aalst, 2003; A. Weijters y Ribeiro, 2011), los algoritmos heurísticos tienen en cuenta las frecuencias de eventos y secuencias al momento de construir el modelo de procesos. La idea principal es que los caminos infrecuentes no deben ser incorporados al modelo (Van der Aalst, 2011). Este enfoque es más robusto ante el ruido y la incompletitud del registro de eventos al considerar métricas basadas en frecuencias (Romero et al., 2016).
- ***Fuzzy Miner*** (Günther y Van Der Aalst, 2007), el enfoque difuso proporciona un conjunto extensible de parámetros para determinar qué actividades y arcos deben incluirse. Además, el enfoque puede construir modelos jerárquicos, es decir, actividades menos frecuentes pueden trasladarse a subprocesos (Van der Aalst, 2011).
- ***Genetic Process Mining*** (de Medeiros et al., 2007), el algoritmo alpha y y las técnicas de PM heurísticas y difusas proporcionan un proceso de manera directa y determinista. Los enfoques evolutivos utilizan un procedimiento iterativo para imitar el proceso de evolución natural. Tales enfoques no son deterministas y dependen de la aleatorización para encontrar nuevas alternativas (Van der Aalst, 2011).

Dentro de las técnicas de verificación de conformidad es posible destacar dos:

- **Verificación de conformidad** (Rozinat y Van der Aalst, 2008), dentro de esta se definen dos dimensiones: el *fitness*, que se refiere a si el registro de eventos puede ser el resultado del proceso modelado y el *appropriateness*, que se refiere a si el modelo es un candidato probable desde un punto de vista estructural y de comportamiento. Para calcular el *fitness*, que es ampliamente utilizado en PM, se utiliza una técnica denominada *token replay*, que consiste en ejecutar las actividades del registro de eventos sobre una red de Petri que representa el modelo de proceso.
- ***Linear Temporal Logic* (LTL) *Checker*** (Van Dongen et al., 2005; van der Aalst et al., 2005), este verifica si los registros de eventos satisfacen alguna fórmula de *Linear Temporal Logic* (LTL), en este caso no se compara un modelo con el registro, sino un conjunto de requisitos descritos por la LTL.

El análisis de gráficas de puntos (*Dotted Chart Analysis*) es incluido dentro de las técnicas de PM al ser implementado como un *plugin* para la herramienta ProM (Song y van der Aalst, 2007), este tipo de gráficas muestra la dispersión de los eventos a lo largo del tiempo trazando un punto para cada evento en un registro de eventos que permite examinar visualmente un registro de eventos y así resaltar algunos patrones interesantes presentes en este. Las gráficas de puntos tienen dos dimensiones ortogonales: tiempo y tipos de componentes. El tiempo se mide a lo largo del eje horizontal del gráfico. Los tipos de componentes (por ejemplo, instancia, creador, tarea, tipo de evento, etc.) se muestran a lo largo del eje vertical (Cairns et al., 2015). En la figura 2.4, se muestra un ejemplo de este tipo de gráficas, la que corresponde al trabajo diario realizado por los alumnos de Moodle, el tamaño de los puntos indica el número de estudiantes realizando la actividad.

El *Social Network Analysis* (SNA) se refiere al conjunto de métodos, técnicas y herramientas de la sociometría destinadas al análisis de la estructura y composición de los vínculos en las redes sociales (Aggarwal y Wang, 2011). En el caso de los MOOCs, es posible utilizar datos de foros para realizar este tipo de análisis y explicar los mecanismos dinámicos de interacción entre los participantes (J. Zhang et al., 2016).

Figura 2.4. Ejemplo de gráfica de puntos. Extraído de Bogarín et al. (2018).

Dentro de las herramientas disponibles para ejecutar las técnicas mencionadas anteriormente, destacamos las siguientes:

- **ProM:** el marco de trabajo de esta herramienta es presentado en Van Dongen et al. (2005), es de código abierto, y tiene diversos *plugins* que implementan todas las técnicas de PM mencionadas ya en este capítulo, además de varias otras técnicas, junto con *plugins* para importación y exportación de modelos y datos, y conversión entre formatos de los modelos y registros de eventos.
- **Disco y Celonis:** son dos herramientas separadas, cuyas sus funcionalidades son similares, enfocadas ambas en el descubrimiento de procesos. Estas herramientas se basan en el algoritmo *fuzzy miner* (Günther y Van Der Aalst, 2007) combinado con algunas características de los algoritmos heurísticos (Van Der Aalst, 2011), en Günther y Rozinat (2012) se presenta el algoritmo de Disco. Sin embargo, en la información pública de ambas herramientas no se cuenta con información suficiente que permita definir cuáles son las diferencias entre los algoritmos de cada una, además de contar con interfaces similares entre sí, donde los parámetros que se pueden variar en ambas herramientas son los mismos. Los algoritmos de dichas herramientas, están diseñados con un foco en manejar procesos complejos y

no estructurados, y que dan como resultado modelos de mapeo de procesos que pueden ser operados y comprendidos por expertos en dominios sin experiencia previa en minería de procesos (Günther y Rozinat, 2012). Finalmente, ambas herramientas comerciales integran un conjunto de métricas y opciones de filtrado para adaptar el registro de eventos a las preguntas específicas y analizar el proceso de forma interactiva. Como se indicará en el capitulo 4, estas herramientas fueron las utilizadas para desarrollar el trabajo que se presenta en el caso de estudio de Maldonado-Mahauad et al. (2018).

- **Softlearn:** es una herramienta específica para EPM, la cual utiliza un algoritmo genético para descubrir itinerarios de aprendizaje completos, de forma que permita a los profesores utilizar esta plataforma para evaluar las actividades de aprendizaje. Además, la plataforma SoftLearn tiene una interfaz gráfica específicamente desarrollado para visualizar tanto las vías de aprendizaje descubiertas por el algoritmo genético y los datos generados durante las actividades de aprendizaje. Puede interactuar con datos de distintos VLE, por medio de *VLE-adapters* específicos para cada entorno de aprendizaje (Barreiros et al., 2014).

2.5. Problemáticas comunes en aplicación de EPM a MOOCs

Existen múltiples problemáticas al momento obtener o generar un registro de eventos que será utilizado en PM, el registro de eventos es de vital importancia, dado que es el punto de partida para aplicar PM, de acuerdo al Process Mining Manifesto, la madurez de un registro de eventos de acuerdo a los datos disponibles, se puede clasificar en cinco niveles (Van Der Aalst et al., 2011):

- ***** **(cinco estrellas):** es el nivel más alto, donde el registro de eventos es de excelente calidad, es decir, confiable y completo, con eventos bien definidos. Los eventos se registran de manera automática, sistemática, confiable y segura. Se toman en cuenta adecuadamente consideraciones acerca de la privacidad y seguridad. Además, los eventos registrados (y sus atributos) tienen una semántica clara.

Esto implica la existencia de una o más ontologías, donde los eventos y sus atributos se refieren a esta ontología.

- ****** (cuatro estrellas):** el registro de eventos es confiable y completo, pero no necesariamente cuenta con una semántica clara. A diferencia de los registros de eventos del nivel ***, se da soporte de forma explícita a nociones tales como instancia de proceso (caso) y actividad.
- ***** (tres estrellas):** los eventos se registran automáticamente, pero no se sigue un enfoque sistemático para ello. A diferencia del nivel **, existe un nivel de garantía que los eventos registrados calzan con la realidad, es decir, el registro de eventos es confiable, pero no necesariamente completo. Aunque se requiere extraer los eventos desde una variedad de tablas y fuentes, se puede asumir que la información es correcta. En general las plataformas MOOC y VLE suelen clasificar en esta categoría, donde por ejemplo es razonable asumir que, si un estudiante responde una evaluación o revisa cierto contenido de un curso, esto se registrará en alguna tabla, y viceversa, pero dependiendo del tipo de actividad realizada, ésta podría ser registrada en tablas distintas.
- **** (dos estrellas):** los eventos se registran automáticamente, como un subproducto de algún sistema de información. La cobertura varía, es decir, no se sigue un enfoque sistemático para decidir qué eventos se registran. Además, es posible pasar por alto el sistema de información. Por lo tanto, podrían faltar eventos o estos podrían no registrarse correctamente.
- *** (una estrella):** es el nivel más bajo, donde los registros de eventos son de mala calidad. Los eventos registrados podrían no corresponder a la realidad y podrían faltar eventos. Esto puede ocurrir en el caso de los registros de eventos creados manualmente.

Teniendo en consideración esta clasificación para la madurez de los registros de eventos, éstos pueden presentar diversos problemas (Van der Aalst, 2011; Cairns et al., 2015; Bogarín et al., 2018), los más relevantes en relación a la utilización de un registro de eventos son:

- **Ruido:** un registro de eventos se considera ruidoso cuando tiene actividades excepcionales que no calzan con la actividad típica del proceso, en el caso de EPM puede ocurrir por razones como el que un estudiante deje su sesión abierta o que haga click de forma errónea en algún lugar de la plataforma.
- **Incompletitud:** esto se refiere a la falta de eventos, y también a que los eventos con los que se cuenta en el registro son pocos como para descubrir los patrones de comportamiento buscados, podría ocurrir tanto por fallos del sistema, o por limitaciones en el diseño de éste, como recordar solo la última interacción con cada actividad disponible en un curso, por ejemplo.
- **Distribución:** la información está distribuida en distintas tablas y fuentes, las que no necesariamente están integradas en un mismo sistema, esto requiere tomar definiciones sobre cómo integrar la información, lo que puede sesgar los resultados.
- **Marca de tiempo:** se requiere que las actividades estén correctamente ordenadas en el tiempo para poder aplicar técnicas de PM, dependiendo del tipo de análisis a realizar, también se pueden necesitar las marcas de tiempo específicas de cada actividad, lo que puede tener complicaciones en un sistema con retardo al registrar la actividad, que guarde solo la fecha o que tenga diferencias de huso horario entre las distintas tablas.
- **Granularidad:** dependiendo de la tabla o la fuente de información consultada, las actividades pueden registrarse con distinto nivel de granularidad, lo que requiere homologarlas, por ejemplo, podrían co-existir registros a nivel de clicks individuales en ciertas actividades, junto con registros solo del resultado final de una actividad completa.
- **Panorámica:** esto se refiere a cuando las instancias del proceso que se desean estudiar, no calzan temporalmente con el registro de eventos considerado, es decir que el caso del proceso comienza antes que el registro de eventos o termina después que este, problema que puede ocurrir si los datos levantados para el registro son filtrados temporalmente.

- **Privacidad:** son muy relevantes las consideraciones sobre la privacidad de los datos, y que los participantes de entornos virtuales de aprendizaje sean conscientes de qué se está haciendo con sus datos, actualmente, legislaciones como la *General Data Protection Regulation* (GDPR) en la Unión Europea exigen informar de antemano a participantes de entornos de aprendizaje en línea sobre lo que ocurrirá con sus datos personales de forma que puedan dar su consentimiento (Leitner et al., 2018), pero desafortunadamente, la anonimización de los datos personales para eludir el problema de privacidad vuelve más difícil el uso de técnicas de *Learning Analytics* en general y no es tan trivial (Khalil y Ebner, 2016), esto suele ser principalmente relevante en el caso de cursos SPOC (*Small Private Open Courses*, que se refiere a cursos de pequeña escala implementados por medio de plataformas MOOC), donde existe un manejo directo de datos de los participantes, a diferencia del caso de cursos masivos, donde las consideraciones de privacidad quedan en manos del proveedor de la plataforma MOOC más que del lado de los investigadores.

En el caso particular de la utilización de técnicas de PM para el estudio de comportamiento en MOOCs, se destacan los siguientes aspectos como los más relevantes a partir de la experiencia del estudio presentado en Maldonado-Mahauad et al. (2018):

- Volumen y complejidad de datos: en particular en el caso de las plataformas MOOC, donde se tienen cursos de miles de estudiantes y en algunos casos estos pueden interactuar libremente y en cualquier orden con las actividades del curso, el volumen de datos puede presentar problemas técnicos y de capacidad de análisis.
- Adaptación al contexto: las distintas plataformas MOOC registran diferentes tipos de actividad, donde además dependiendo del diseño del curso pueden usarse solo ciertos tipos de actividad en particular, lo que impide generar procedimientos de análisis en detalle generalizables para todos los cursos, favoreciendo el uso de metodologías flexibles.

- **Determinar el enfoque de investigación:** la actividad en cursos MOOC suele obedecer a procesos no estructurados, y que puede ser relevante analizar desde distintos niveles de detalle, como puede ser a nivel del curso en general, o a cómo trabajan los estudiantes en una sesión de estudio en particular, por lo que tener bien definido el enfoque de investigación es muy importante, ya que la instancia del proceso no siempre puede ser determinada explícitamente de los datos, sino que obedece a definiciones del mismo enfoque de investigación.
- **Identificar modelos teóricos educativos:** es de suma importancia contar con modelos teóricos educativos sobre los cuales comparar los resultados obtenidos, ya que, sin estos, resulta difícil darles sentido y utilidad a las conclusiones del análisis. Por otro lado, contar con un marco teórico bien estructurado facilita la iteración en el análisis y sirve de guía para las definiciones que se deben tomar al aplicar técnicas de PM.

3. PROPUESTA METODOLÓGICA DE MINERÍA DE PROCESOS PARA ANÁLISIS DE DATOS EDUCACIONALES EN CURSOS MOOC

En este capítulo se presenta una propuesta metodológica para el uso de PM en el análisis de datos educacionales en MOOCs, la cual está basada en la metodología PM^2, por lo que primero se presentan los aspectos generales de dicha metodología, y luego las adaptaciones realizadas para el uso particular en un contexto MOOC.

3.1. Metodología PM^2

Dentro de la minería de datos existen múltiples esfuerzos por crear metodologías de proyecto, dentro de estas destacan CRISP-DM y SEMMA, las que sin embargo son de muy alto nivel y proporcionan poca orientación a actividades especifícas de PM (Van der Aalst, 2011). Debido a esto en el campo de PM se han propuesto diversas metodologías más específicas a sus actividades propias, siendo las primeras de éstas Process Diagnostics Method (PDM) (Bozkaya et al., 2009), la que ha sido aplicada no solo a negocios, sino que también en el área de la salud (Rebuge y Ferreira, 2012), y también el modelo de ciclo de vida L* (Van der Aalst, 2011). PDM está diseñado para proporcionar rápidamente una visión general de un proceso, mientras que L* cubre muchos aspectos diferentes de la minería de procesos y aborda temas más amplios como la mejora de los procesos y el soporte operacional (van Eck et al., 2015).

Sin embargo estas metodologías presentan diversas limitaciones. Por un lado, PDM, cubre solo un pequeño número de técnicas de PM (van Eck et al., 2015), enfatizando en cómo evitar el uso de conocimiento experto durante el análisis (Bozkaya et al., 2009), lo que la hace menos aplicable a proyectos grandes y complejos (Suriadi et al., 2013). L* cubre técnicas más diversas, pero está diseñado principalmente para el análisis de procesos estructurados y tiene como objetivo descubrir único modelo de proceso integrado (Van der Aalst, 2011). Además de lo ya mencionado, ni PDM ni L* fomentan explíticamente el

análisis iterativo, y en el caso de ambas metodologías, éstas pueden ser complementadas con directrices prácticas adicionales para ayudar a los profesionales sin experiencia a superar los retos comunes (van Eck et al., 2015).

Teniendo esto en consideración es que se crea la metodología PM^2 (van Eck et al., 2015), la cual está diseñada para apoyar proyectos destinados a mejorar el rendimiento de los procesos o el cumplimiento de normas y reglamentos. PM^2 cubre una amplia gama de técnicas de minería de procesos y otras técnicas de análisis, y es adecuado para el análisis tanto de procesos estructurados como no estructurados. En un contexto MOOC, dependiendo del enfoque de la investigación el proceso a analizar puede ser estructurado, como es el caso del recorrido de los estudiantes por un curso completo (Mukala, Buijs, Leemans, y van der Aalst, 2015), sin embargo, para otros enfoques de análisis que no contemplen directamente la estructura del curso, ya sea el curso completo, o sea un capítulo o semana de este, el proceso a estudiar podría ser no estructurado, por lo que PM^2 ofrece un marco de trabajo más adecuado para el estudio de datos educacionales de cursos MOOC, que las metodologías anteriormente mencionadas.

De acuerdo a lo presentado por van Eck et al. (2015), la metodología PM^2 consiste de seis etapas que se relacionan con diferentes objetos de entrada y salida de los siguientes tipos: objetos relacionados con los objetivos, objetos de datos y modelos. Los cuatro objetos relacionados con los objetivos son (1) preguntas de investigación derivadas de las metas del proyecto, las cuales son respondidas por (2) desempeño y (3) resultados de cumplimiento, lo que conduce a (4) ideas de mejora para lograr los objetivos. Los objetos de datos denotan las tres representaciones diferentes de datos relacionados con el proceso: (1) los sistemas de información contienen datos vivos del proceso en diversas formas, que pueden ser extraídos y vinculados a eventos discretos para formar (2) datos de eventos. Los datos de eventos pueden ser transformados en (3) registros de eventos mediante la definición del caso y de las clases de eventos. Se consideran dos tipos de modelos: (1) modelos de proceso y (2) modelos analíticos. Los modelos de proceso describen la ordenación de actividades en un proceso, siendo posible ampliarlo con información adicional,

por ejemplo, restricciones temporales, uso de recursos o uso de datos. Además, también se consideran las reglas de negocio (en abstracto) como modelos de procesos que definen formalmente las restricciones con respecto a la ejecución de los procesos de negocio. Los modelos analíticos son cualquier otro tipo de modelos que puedan aportar información sobre el proceso, por ejemplo, árboles de decisión.

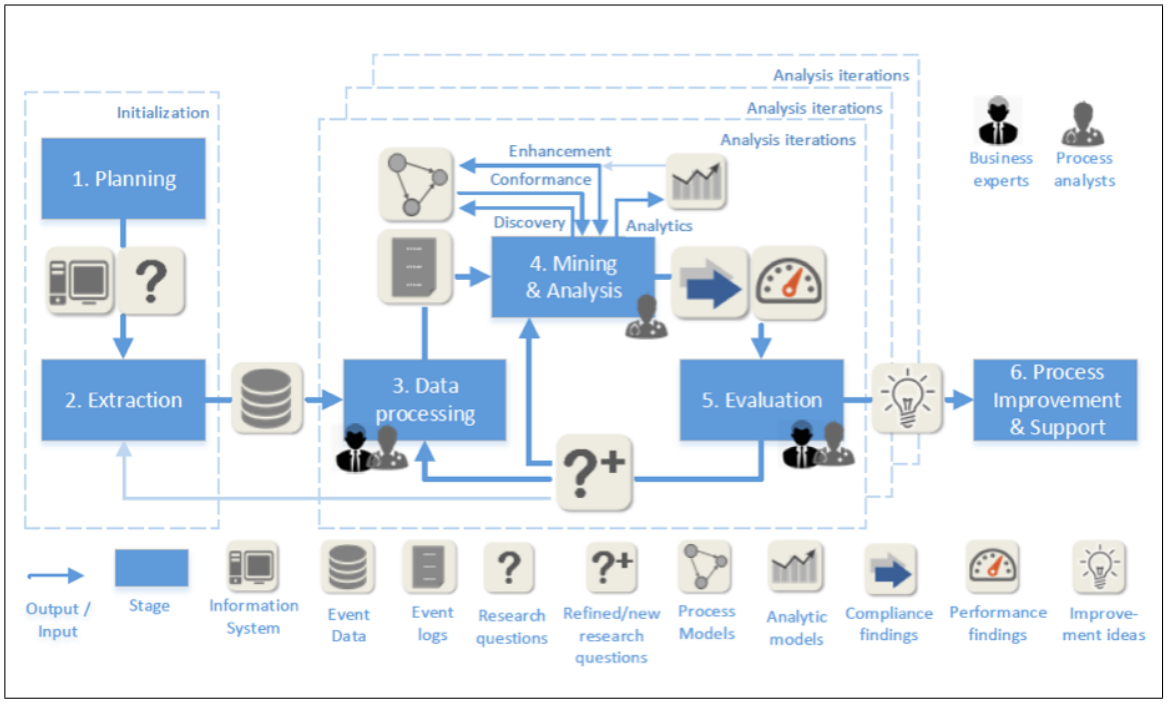


Figura 3.1. Etapas de la metodología PM^2. Extraída de van Eck et al. (2015).

Como se puede apreciar en la figura 3.1, las dos primeras etapas de la metodología son (1) la planificación y (2) la extracción, durante las cuales se definen las preguntas iniciales de investigación y se extraen los datos del evento. Después de la primera dos etapas, se realizan una o más iteraciones de análisis, posiblemente en paralelo. En general, cada iteración de análisis ejecuta las siguientes etapas una o más veces: (3) procesamiento de datos (4) minería y análisis, y (5) evaluación. Cada iteración del análisis se centra en responder a una pregunta de investigación específica mediante la aplicación de actividades relacionadas con la minería de procesos y la evaluación de los modelos de procesos descubiertos y otros hallazgos. Esta iteración puede puede tardar desde minutos hasta días

en completarse, dependiendo principalmente de la complejidad de la minería y análisis. Si los resultados son satisfactorios, estos pueden utilizarse para (6) mejora de procesos y soporte.

3.2. Adaptaciones realizadas en propuesta metodológica de PM aplicado a contexto MOOC

La metodología PM^2, al igual que las demás metodologías de minería de datos y minería de procesos mencionadas, está pensada en primera instancia para un contexto de procesos de negocios, lo que vuelve necesario adaptar las fases y actividades descritas de la metodología original, para hacerlas adecuadas en un contexto MOOC.

En esta propuesta metodológica que adapta PM^2 al contexto MOOC, consideraremos cuatro etapas (Figura 3.2): (1) Extracción - los datos se extraen de las bases de datos del MOOC; (2) Generación del log de eventos - la información de la tabla se modela en términos de registros de eventos, definiendo lo que es un caso (ejecución de un proceso), actividades (pasos del proceso), y un orden temporal de las actividades; (3) Descubrimiento del modelo - se aplican algoritmos de minería de procesos para el descubrimiento del modelo y que describe el comportamiento observado; y (4) Análisis del modelo – el modelo de procesos descubierto es analizado a fin de comprender el comportamiento observado.

En relación a la etapa 1 de la metodología PM^2 original, correspondiente a la planificación, es decir: la definición de procesos a estudiar, formulación de preguntas de investigación y creación del equipo del proyecto, no se considera necesaria una explicación de forma detallada como parte del trabajo de tesis, dado que se asume la realización de estas actividades como parte del trabajo investigativo en general, por lo que la metodología de trabajo no requiere directrices específicas para el contexto MOOC. Sin embargo, las definiciones de los procesos a estudiar y las preguntas de investigación del proyecto, se asumen como variables de entrada para las etapas descritas en esta propuesta metodológica.

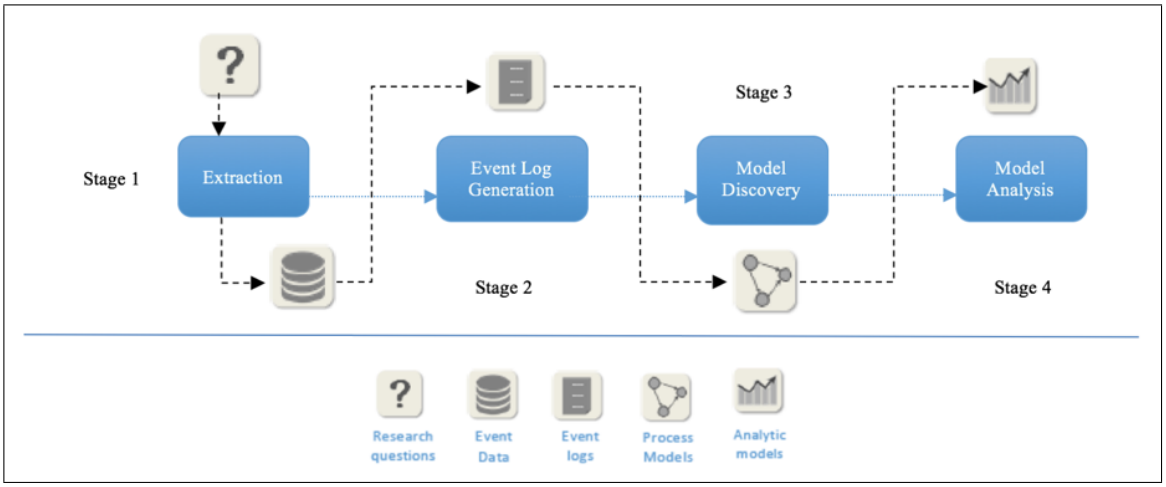


Figura 3.2. Etapas de la propuesta metodológica de PM para análisis de datos educacionales en en cursos MOOC, basada en la metodología PM^2. Extraída de Maldonado-Mahauad et al. (2018), adaptada de van Eck et al. (2015).

Además, dadas las limitaciones prácticas de trabajar en el contexto MOOC, no siempre resulta posible aplicar directamente la etapa 6 de la metodología PM^2 original, de mejora de procesos y soporte, dado que las plataformas MOOC desde donde se obtienen los datos, en general son de administración externa a los equipos de investigación que las estudian, salvo el caso de Open edX, aun así, no resulta factible modificar la estructura en que la plataforma presenta los cursos, la que es una de las principales condicionantes del proceso a estudiar, sino que esa es tarea de las empresas que han creado las plataformas MOOC, como Coursera, quienes pueden utilizar como insumo los resultados de este tipo de investigaciones. De todas formas es posible entregar los resultados obtenidos de aplicar la metodología a profesores y diseñadores instruccionales, de forma que puedan aplicar cambios sobre el material y/o la estructura de versiones futuras del curso estudiado.

3.2.1. Fase 1: Extracción de datos

La etapa de extracción tiene como objetivo extraer datos de eventos y, opcionalmente en el caso que estén disponibles, modelos de proceso. Los insumos para esta etapa son las preguntas de investigación y los sistemas de información que soportan la ejecución de

los procesos seleccionados del estudio. Los resultados de esta etapa son datos de evento, es decir, una colección de eventos sin noción de caso o clases de evento predefinidas, y posiblemente modelos de proceso (van Eck et al., 2015).

La metodología PM^2 propone 3 actividades como parte de esta etapa: (1) determinación de enfoque, (2) extracción de datos de eventos, y (3) transferencia de conocimiento del proceso. A continuación se explicarán estas actividades y cómo aplicarlas en el contexto MOOC:

(1) **Determinación de enfoque:** esta actividad implica determinar el enfoque de la extracción de datos, en función de la cual deben crearse los datos del evento. Esto debe tener distintas consideraciones, dentro de las que destacan la granularidad de los datos, el período de datos a extraer, los atributos de datos a considerar y por medio de qué correlaciones serán extraídos los datos. En particular para los MOOC, es importante definir la granularidad de los datos a usar, en algunas plataformas, los datos pueden ser obtenidos hasta el detalle de clicks realizados por el usuario, en datos tipo clickstream (Brinton & Chiang, 2015), dependiendo del enfoque de la investigación, puede bastar con una base de datos que incluya la lista de objetos de aprendizaje del curso y cuándo se interactuó con estos.

(2) **Extracción de datos de eventos:** estando determinado el enfoque de la extracción, los datos de evento pueden crearse recogiendo los datos a partir de las bases de datos y clickstreams disponibles, para luego unirlos en una única colección de eventos, por ejemplo, una tabla en la que cada entrada representa un evento.

(3) **Transferencia de conocimiento del proceso:** esta actividad se puede desarrollar en paralelo a la creación de datos de eventos, consiste tanto en el levantamiento de conocimiento teórico sobre el proceso a analizar, como por ejemplo, teorías de autorregulación del aprendizaje relevantes para explicar la actividad desarrollada en el curso (Bannert, Reimann & Sonnenberg, 2014; Kizilcec, Pérez-Sanagustín & Maldonado, 2017; Maldonado et al., 2017), o la estructura del curso como base para el modelo de procesos (Mukala, Buijs & Van Der Aalst, 2015). Contar con

este conocimiento del proceso es indispensable para realizar un procesamiento de datos efectivo, pues sitúa el análisis posterior.

PM^2 considera una división explícita en dos etapas de la extracción de datos de eventos y la creación y procesamiento del registros de eventos, a diferencia de otras metodologías, esto porque la extracción de datos de eventos consume mucho tiempo y se repite con menos frecuencia que las actividades de procesamiento de datos como el filtrado. Además de ser posible el crear diferentes vistas sobre los mismos datos de evento, lo que da lugar a diferentes registros de eventos (Van Eck, Lu, Leemans & Van der Aalst, 2015).

3.2.2. Fase 2: Generación de registro de eventos

En esta etapa se busca crear registros de eventos como vistas diferentes de los datos de eventos obtenidos y procesar registros de eventos de tal manera que sean óptimos para las siguientes fases, siendo una de las etapas más relevantes en la metodología, puesto que el registro de eventos es la base para los modelos de proceso a descubrir y los análisis a realizar posteriormente. Además de los datos de evento como entrada principal, también se pueden utilizar modelos de proceso como entrada para filtrar los datos de evento. Los resultados son registros de eventos que se utilizan en la etapa de extracción y análisis. Las actividades a realizar en esta etapa son: (1) creación de vistas, (2) combinación de eventos, y (3) enriquecimiento de registros, las que son descritas a continuación:

(1) **Creación de vistas:** esta actividad implica definir los tipos de eventos a considerar y el caso del proceso. Dependiendo del nivel de análisis y detalle deseado, algunos ejemplos de casos posibles a considerar en un MOOC serían: curso completo, capítulo del curso, semana del curso, sesión de trabajo, intervalo entre evaluaciones, un vídeo o evaluación en particular, entre otros. Teniendo en consideración el caso definido los tipos de eventos deben definirse de forma que sea comparable una instancia del proceso con otra, es decir, si el caso fuera una sesión de trabajo en el MOOC, los tipos de eventos podrían ser lecturas, vídeos y evaluaciones, sin considerar en particular cuál lectura, vídeo o evaluación fue la realizada, de forma

que sea comparable el proceso desarrollado al inicio del curso, con el realizado posteriormente.

(2) **Combinación de eventos:** La combinación de eventos puede ayudar a reducir la complejidad y mejorar la estructura del modelo a minar (Bose y Van der Aalst, 2009). En van Eck et al. (2015) se indica que existen dos tipos de combinación posibles: is-a y part-of.

La primera considera diferentes tipos de eventos pertenecientes a una clase de evento equivalente, pero más general, por lo que se mantiene el número de eventos, un ejemplo de esto son las evaluaciones formativas y las evaluaciones sumativas, donde ambas podrías ser denominadas solo como evaluaciones. Por otro lado el segundo tipo de combinación fusiona múltiples eventos en uno mayor, por ejemplo, si un curso al final de un capítulo tuviese múltiples actividades de evaluación, el ejecutar estas actividades de evaluación, en lugar de considerarse como eventos por separado, podrían considerarse como solo un evento de evaluaciones de final de capítulo.

(3) **Enriquecimiento de registros:** el registro puede ser enriquecido tanto con datos derivados del mismo registro, como con datos externos. Un ejemplo de lo primero sería el tiempo total de una sesión de trabajo calculado desde el mismo registro, mientras que un ejemplo de lo segundo podría ser incluir información sobre si el estudiante aprobó el curso, o el resultado de alguna medición externa sobre éste.

3.2.3. Fase 3: Descubrimiento del modelo de proceso

Esta fase consiste en la aplicación de algún algoritmo de descubrimiento de procesos, para generar un modelo de procesos real como salida, teniendo como entrada para esta etapa el registro de eventos creado anteriormente, en este caso estamos considerando una subdivisión de la etapa 4 de PM^2, entre esta fase de la propuesta metodológica y la siguiente.

El descubrimiento del modelo de proceso puede realizarse con alguna de las herramientas mencionadas durante el capítulo 2. Para el caso de procesos no estructurados y estudios exploratorios, una alternativa es trabajar con herramientas como Disco o Celonis, que usan el algoritmo *fuzzy miner*, donde es posible visualizar el modelo de proceso con distintos niveles de complejidad de acuerdo a los parámetros utilizados, y también es posible la aplicación de filtros sobre el registro dentro de la misma herramienta.

En caso de ser necesario reducir la complejidad o centrar el análisis en una parte específica del conjunto de datos, es posible utilizar técnicas habituales del filtrado sobre el registro de eventos, como *slice and dice* (filtrar por atributos), filtrado basado en varianza y filtrado basado en cumplimiento, esto permitirá obtener múltiples modelos de proceso para cada partición, los que podrían ser más sencillos que el modelo de proceso minado del registro completo (Rebuge y Ferreira, 2012). Para el caso del *slice and dice*, algunos ejemplos serían filtrar solo alumnos que aprueban el curso, o filtrar solo las actividades de vídeo donde el estudiante vio más de cierto tiempo mínimo el vídeo. En el filtrado basado en varianza, se busca agrupar trazas similares, ya sea aplicando técnicas de clusterización o considerando ciertas transiciones relevantes entre los eventos. Mientras que el filtrado basado en cumplimiento consiste en eliminar trazas o eventos que no cumplan una regla en particular, o no calcen con un modelo de proceso establecido previamente.

3.2.4. Fase 4: Análisis del modelo

Esta fase corresponde a la segunda parte de la etapa 4 de PM^2, sumado a la etapa 5 de evaluación, donde contando ya con el modelo de proceso general, o modelos específicos para distintas particiones del registro de eventos, es posible desarrollar distintas actividades de análisis sobre estos, las actividades a considerar en esta fase son: (1) verificación de conformidad y mejoramiento, (2) uso de analíticas y (3) evaluación. Con esto se busca obtener respuesta a las preguntas de investigación, o la iteración sobre mejoras al proceso y nuevas preguntas de investigación que sea posible responder con los datos.

(1) **Verificación de conformidad y mejoramiento:** en caso de contar con un modelo de proceso teórico sobre el cual se quiere comparar el modelo de proceso real, la herramienta ProM provee de plugins necesarios para realizar verificación de conformidad y otras técnicas de PM posteriores al descubrimiento.

(2) **Uso de analíticas:** por otro lado, además de las técnicas propias de PM, es posible ampliar el análisis por medio de otras técnicas de minería de datos, como la clasificación de los estudiantes en base a información obtenida de los modelos de procesos, o la visualización de tiempos o números de eventos asociados a cada caso, de forma que se pueda responder de mejor forma a las preguntas de investigación.

(3) **Evaluación:** de acuerdo a PM^2, la evaluación debe contar de dos partes, el diagnóstico y la verificación y validación. En el diagnóstico se busca interpretar correctamente los resultados, distinguir los resultados esperados de aquellos novedosos o inusuales, y refinar las preguntas de investigación para iteraciones posteriores. Mientras que por otro lado, en la verificación y validación, se debe investigar la exactitud de los hallazgos inesperados, tanto comparando resultados obtenidos con los datos originales (verificación), como comparando los resultados con afirmaciones de las partes interesadas en el proceso (validación), lo que implica levantar información desde estudiantes y/o profesores de los curso para ello. La verificación y validación puede ayudar a identificar las causas subyacentes y a diseñar ideas para una posible mejoras en los procesos.

4. CASO DE ESTUDIO: APLICACIÓN DE LA PROPUESTA METODOLÓGICA PARA EL ANÁLISIS DE LA ACTIVIDAD DE ESTUDIANTES DE CURSOS MOOC EN LA PLATAFORMA COURSERA

En este capítulo se presenta un caso de estudio en donde es aplicada la propuesta metodológica del capítulo 3, esta investigación fue presentada en el artículo *Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses* de Maldonado-Mahauad et al. (2018), publicado en *Computers in Human Behavior*, donde el autor de esta tesis contribuye como cientista de datos en la investigación. Esta aplicación sirve para validar la aplicación de la metodología propuesto y para ejemplificar su uso. El objetivo de este capítulo es validar las dos preguntas de investigación iniciales:

- RQ1. ¿Es posible adaptar y proponer una metodología basada en técnicas de PM para estudiar el comportamiento de los estudiantes a partir de datos de una plataforma MOOC?
- RQ2. ¿Qué variaciones debo realizar en la metodología de PM para extraer patrones de comportamiento y perfiles de estudiante a partir de datos de varios MOOCs?

En relación a la RQ1, el caso de estudio aquí presentado analiza un caso concreto para entender la autorregulación de los estudiantes en un MOOC. Esto nos permitierá entender si la metodología propuesta aplica para un caso real de análisis.

En relación a la RQ2, el aplicar la metodología para un caso real nos permitirá entender qué variaciones son necesarias para un contexto de análisis real del que se deben extraer conclusiones.

4.1. Contexto del caso de estudio y objetivo de análisis

4.1.1. Objetivos de investigación del caso de estudio

En esta investigación, se aplica la metodología propuesta para comprender los patrones de comportamiento en plataformas MOOC a gran escala y la relación que estos pueden

tener con el éxito en el curso y los niveles auto-reportados de autorregulación (SRL, a partir de ahora, de sus siglas en inglés), ampliando hallazgos anteriores en el área. Esto busca responder las siguientes preguntas, particulares de este caso de estudio:

1. ¿Es posible utilizar minería de procesos para identificar las secuencias de interacción más frecuentes de los estudiantes que reflejen estrategias de aprendizaje específicas? En caso afirmativo, ¿qué tipos de interacciones pueden identificarse?
2. ¿En qué se diferencian las secuencias de interacción de los estudiantes con diferente rendimiento académico?
3. ¿En qué se diferencian las secuencias de interacción entre los estudiantes con diferentes perfiles de auto-regulación?

Para resolver estas preguntas de investigación, a continuación, describiremos cómo fue aplicada la metodología, tanto en cómo se adaptaron las fases de la metodología a la plataforma Coursera, como los elementos adicionales empleados en el análisis.

La investigación original contempla una cuarta pregunta de investigación referida a cómo se relacionan las secuencias de interacción identificadas con las estrategias de SRL descritas en la literatura, esta no será tratada como parte del caso de estudio, dado que no puede ser respondida solo desde la propuesta metodológica de PM.

4.1.2. Muestra

La muestra final del estudio está compuesta por N = 3458 estudiantes inscritos en tres diferentes MOOCs de la plataforma Coursera, ofrecidos por la Pontificia Universidad Católica de Chile. Esta muestra es un subconjunto de 4871 respuestas recibidas de estudiantes que contestaron de forma voluntaria un cuestionario de auto-reporte de SRL (aprendizaje auto-regulado, de su sigla en inglés) e intenciones para matricularse en el curso, este cuestionario fue enviado a la totalidad de los 54935 estudiantes que se encontraban registrados en los cursos a la fecha de extracción de los datos. La recolección de datos del cuestionario ocurrió entre abril y diciembre de 2015, extrayéndose los datos de actividad

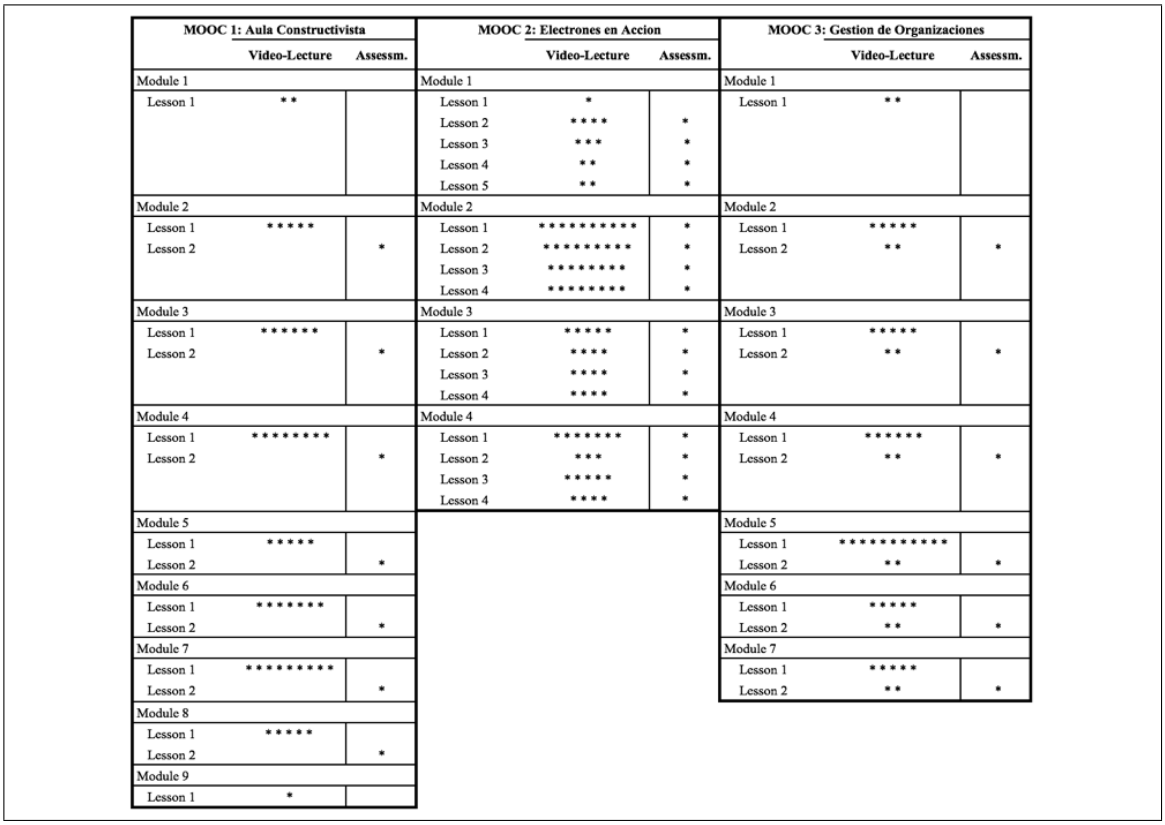
en la plataforma al finalizar este periodo. De las respuestas recibidas, fueron excluidas 1413 por alguna de las siguientes razones: (1) era una respuesta nueva recibida desde un estudiante que ya había contestado anteriormente el cuestionario (N = 733), (2) cuestionarios vacíos sin respuestas (N = 133), y (3) los datos de los cuestionarios no pudieron vincularse con los datos registrados en la plataforma (N = 547).

El público objetivo de los tres cursos eran estudiantes de secundaria, estudiantes universitarios y profesionales de las industrias relacionados con cada asignatura. En base a los datos demográficos capturados durante el proceso de registro en la plataforma, la edad promedio fue de 32 años (SD $= 11{,}07$). Un cuarto de los estudiantes eran mujeres y el 88 % cuenta con una licenciatura o un título superior (14 % de maestría o doctorado).

Los cursos se impartieron en español, y corresponden a los siguientes:

1. Hacia una práctica constructivista en el aula. Curso de educación, con n = 497 participantes en la muestra de estudio final.
2. Electrones en Acción. Curso de ingeniería, con n = 2.035 participantes.
3. Gestión de Organizaciones Efectivas. Curso de gestión, con n = 926 participantes.

Los cursos se organizan en módulos, cada uno de los cuales consta de varias lecciones y estas a su vez contienen múltiples vídeos (vídeo-lecciones). La Figura 4.1 ilustra la estructura de cada curso, cada curso se estructura en módulos, cada módulo se compone de lecciones, y cada lección se compone de video-lecturas y evaluaciones, los "*" representan una video-lectura o una actividad de evaluación en cada lección. La tabla 4.1 muestra el número de alumnos matriculados, la tasa de aprobación, los módulos, las lecciones, las video-lecciones y las actividades de evaluación para cada curso. Los cursos siguieron un formato "bajo demanda" en el que los materiales del curso estaban disponibles de una sola vez, sin plazos específicos predefinidos.



MOOC 1: Aula Constructivista		
	Video-Lecture	Assessm.
Module 1		
Lesson 1	**	
Module 2		
Lesson 1	*****	
Lesson 2		*
Module 3		
Lesson 1	******	
Lesson 2		*
Module 4		
Lesson 1	********	
Lesson 2		*
Module 5		
Lesson 1	*****	
Lesson 2		*
Module 6		
Lesson 1	*******	
Lesson 2		*
Module 7		
Lesson 1	*********	
Lesson 2		*
Module 8		
Lesson 1	*****	
Lesson 2		*
Module 9		
Lesson 1	*	

MOOC 2: Electrones en Accion		
	Video-Lecture	Assessm.
Module 1		
Lesson 1	*	
Lesson 2	****	*
Lesson 3	***	*
Lesson 4	**	*
Lesson 5	**	*
Module 2		
Lesson 1	**********	*
Lesson 2	*********	*
Lesson 3	********	*
Lesson 4	********	*
Module 3		
Lesson 1	*****	*
Lesson 2	****	*
Lesson 3	****	*
Lesson 4	****	*
Module 4		
Lesson 1	*******	*
Lesson 2	***	*
Lesson 3	*****	*
Lesson 4	****	*

MOOC 3: Gestion de Organizaciones		
	Video-Lecture	Assessm.
Module 1		
Lesson 1	**	
Module 2		
Lesson 1	*****	
Lesson 2	**	*
Module 3		
Lesson 1	*****	
Lesson 2	**	*
Module 4		
Lesson 1	******	
Lesson 2	**	*
Module 5		
Lesson 1	**********	
Lesson 2	**	*
Module 6		
Lesson 1	*****	
Lesson 2	**	*
Module 7		
Lesson 1	*****	
Lesson 2	**	*

Figura 4.1. Estructura de cada MOOC.

Tabla 4.1. Visión general de la estructura de los 3 MOOCs

	MOOC 1	**MOOC 2**	**MOOC 3**
	(n = 497)	**(n = 2,035)**	**(n = 926)**
Matriculados	18,653	25,706	10,576
Tasa de Aprobación	1.40 %	8.40 %	11.40 %
Módulos	9	4	7
Lecciones	9	17	13
Video-lecciones	48	83	51
Evaluaciones	7	16	6

4.1.3. Instrumentos de medición aplicados

De forma complementaria a los datos de actividad en el curso, los estudiantes de los tres MOOCs completaron un cuestionario opcional al comienzo del curso. Este cuestionario incluía preguntas relacionadas con la demografía de los estudiantes (edad, género, educación) y sus intenciones en el curso (por ejemplo, ver todas las video-lecciones o sólo algunas de ellas). Además, el cuestionario incluyó la escala de Intenciones para Matricularse en Aprendizaje en Línea (OLEI en inglés) (Kizilcec y Schneider, 2015) traducida al español, y una medida de SRL que se utilizó en investigaciones previas con MOOCs (Kizilcec et al., 2017). La medida de SRL se compuso de 24 ítems relacionados con seis estrategias SRL y que fue originalmente adaptada de múltiples instrumentos establecidos en la literatura (Littlejohn y Milligan, 2015; Barnard et al., 2008; Pintrich et al., 1991; Warr y Downing, 2000; Rigotti et al., 2008). Los estudiantes valoraron los ítems utilizando una escala tipo Likert de 5 puntos (codificada de 0 a 4). Las seis estrategias de SRL que fueron evaluadas son: establecimiento de objetivos (4 declaraciones), planificación estratégica (4), autoevaluación (3), planificación estratégica (6), elaboración (3) y búsqueda de ayuda (4). Un ejemplo de un ítem es: ”Leí más allá de los materiales básicos del MOOC para mejorar mi comprensión”. Para cada estrategia, la puntuación individual se calculó promediando las calificaciones de ítems correspondientes a dicha estrategia.

La fiabilidad del cuestionario se obtuvo siguiendo el mismo procedimiento que en trabajos anteriores (Kizilcec et al., 2017). Para ello se calculó el alpha de Cronbach por cada subescala y total con los datos levantados en esta muestra, donde se observó que la medida SRL mostró una alta fiabilidad para todas las subescalas de estrategia, entregando valores del alpha de Cronbach de al menos 0,70, que se considera generalmente aceptable (Peterson, 1994). La composición total de la SRL (que es un índice de las seis subescalas), obtuvo una fiabilidad muy alta (alpha = 0,91). La tabla 4.2 presenta estadísticas descriptivas para cada estrategia de SRL, así como los coeficientes de correlación de Pearson entre las estrategias. En el anexo A se incluye un script de Python utilizado para esta validación del alpha de Cronbach.

Tabla 4.2. Estadísticas descriptivas para cada estrategia de SRL: media y desviación estándar, alpha de Cronbach y los coeficientes de correlación de Pearson entre las estrategias, y la composición total de la SRL.

Estrategia	**M (SD)**	α	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**
1. Goal Setting	3.02 (0.75)	.86	.70	.46	.57	.46	.29	.78
2. Strategic Planning	3.11 (0.64)	.73		.60	.65	.58	.31	.84
3. Self-evaluation	3.28 (0.65)	.79			.62	.60	.24	.73
4. Task Strategies	3.10 (0.62)	.78				.72	.34	.87
5. Elaboration	3.31 (0.63)	.76					.32	.77
6. Help Seeking	2.62 (0.78)	.75						.58
7. SRL Composite	3.06 (0.52)	.91						

4.2. Propuesta metodológica aplicada

4.2.1. Fase 1: Extracción de datos

En esta etapa, se extrajeron las trazas de datos desde la base de datos de Coursera para estudiar las secuencias de interacción de los estudiantes en el MOOC. Coursera es una plataforma que mantiene un registro de casi todos los detalles de las interacciones de los estudiantes. Estos datos "en bruto" se organizan por medio de tres categorías: datos generales, datos de foros y datos personales. Comprende 86 tablas de información. Para el propósito de este estudio, hemos limitado nuestro análisis seleccionando sólo las tablas (13) que contienen información relevante sobre el comportamiento de los estudiantes. Los conjuntos de datos extraídos incluyen información del curso, contenido del curso, progreso del curso, evaluaciones, calificaciones del curso y datos demográficos de los estudiantes (basados en cuestionarios realizados a los estudiantes).

El contenido de estas 13 tablas utilizadas es el siguiente:

1. Users – contiene información sobre los usuarios de Coursera (ID).
2. Courses – contiene el nombre del curso y un ID.

3. Course Grades – contiene el estado de aprobación/reprobación de un estudiante en el curso (la codificación es que 0 significa que inició el curso, pero no lo ha completado).
4. Course Memberships – contiene información acerca del rol del usuario dentro del curso (no inscrito, navegando, estudiante, mentor, etc.).
5. Course Modules – contiene información relacionada a los módulos del curso y su orden.
6. Course Lessons – contiene información relacionada a las lecturas del curso y su orden en cada módulo.
7. Course Items – para cada item del curso contiene información relacionada a la lectura correspondiente, la descripción del item, su tipo y su orden dentro de la lectura.
8. Course Items Types - contiene información sobre la descripción del item, su categoría (por ejemplo: *quiz*, *peer review*, etc.) y si éste es evaluado con una nota.
9. Course Item Assessment – contiene información acerca de las evaluaciones del curso.
10. Course Item Grades – contiene información relacionada a la nota de un item del curso, y su aprobación (por ejemplo: 0 significa no aprobado, 1 significa aprobado).
11. Course Progress – contiene información acerca del progreso del estudiante en el curso y sus items, en este caso cada interacción del usuario con un item genera una fila registrando el momento de la interacción y si el item fue completado o no, por lo que será la principal tabla a considerar para construir el log de eventos.
12. Course Progress State – contiene información acerca del estado de avance de un item (por ejemplo: 1 indica iniciado, 2 indica completado).
13. Course Item Passing State – contiene un identificador y descripción del estado de aprobación de cada item (por ejemplo: 0 significa no aprobado, 1 significa aprobado).

Dentro de las tablas existentes, hay otras que registran distintos tipos de actividad, como lo son la actividad en foros de discusión, los que no fueron considerados dado que la actividad era casi inexistente en dichos foros (menos de 0.1 interacciones en foro por estudiante en promedio). Tampoco fueron consideradas las tablas referentes a actividades que no existen en los cursos estudiados, como lo son las evaluaciones entre pares y las evaluaciones de programación. Al momento de realizar el análisis de los datos, no se contaba con datos tipo *clickstream* de estos cursos, por lo que tampoco fueron considerados.

El enfoque para la extracción de los datos fue de considerar las actividades completas, es decir estudiar cada vídeo-lectura y cada evaluación de forma íntegra, sin considerar el detalle de las interacciones con el vídeo o el detalle de cada pregunta de la evaluación, pero si considerando el resultado final de la actividad, es decir, sí fue completada o no.

4.2.2. Fase 2: Generación de registro de eventos

En esta etapa, se define el registro de eventos que se procesará con un algoritmo de minería de procesos. El registro de eventos contiene información sobre las interacciones realizadas por los estudiantes con las vídeo-lecturas y evaluaciones disponibles en cada curso. Dentro de la información complementaria que fue incorporada al registro de eventos, se encuentra el puntaje del cuestionario de auto-reporte de SRL y el estado de aprobación del curso del estudiante.

Para generar el registro de eventos, primero se definieron dos conceptos importantes para referirse a las trazas de datos registradas en la base de datos de Coursera. Específicamente, se definieron los siguientes conceptos de interacción y sesión como sigue:

- Una **interacción** es una acción registrada en la traza de datos de Coursera que registra la interacción de un estudiante con un objeto del MOOC (por ejemplo, video-lecturas y evaluaciones). Se definieron seis tipos de interacciones en función de los objetos con los que interactúan los estudiantes: iniciar una video-lectura, completar una video-lectura, revisar una video-lectura ya terminada, intentar una

evaluación, pasar una evaluación y revisar una evaluación ya aprobada, estas son descritas en la tabla 4.3. Además de estas interacciones, también se incluyó una etiqueta para identificar la primera y la última interacción del estudiante con el curso, a la que se denominó como sesión de inicio y sesión final, respectivamente. Dentro del marco de trabajo de PM, las distintas interacciones corresponderían a las clases de evento.

- Una **sesión** es un período de tiempo en el que el registro de datos de Coursera registra la actividad continua de un estudiante dentro del curso, con intervalos de inactividad no superiores a 45 minutos. Esta definición de sesión fue adoptada a partir de los trabajos anteriores de Kovanović et al. (2015) y Liu et al. (2015). Según está definida en esta investigación, la sesión correspondería al caso o instancia del proceso.

Tabla 4.3. Definición de 6 tipos de interacciones con los materiales del curso que caracterizan el comportamiento continuo del estudiante. Las codificaciones de los nombres de las interacciones se han mantenido en inglés.

Interacción	**Definición**
(1) Video-Lecture begin	Ver una video-lectura sin completarla. La video-lectura no se ha completado previamente.
(2) Video-Lecture complete	Ver una video-lectura en su totalidad en el primer intento.
(3) Video-Lecture review	Regresar a una video-lectura que el estudiante había visto previamente en su totalidad (no necesariamente en el primer intento).
(4) Assessment try	Intento sin éxito de resolver una evaluación.
(5) Assessment pass	Intento exitoso de resolver una evaluación.
(6) Assessment review	Volver a una evaluación que previamente fue completada con éxito (no necesariamente en el primer intento).

Para las interacciones de cada usuario, fueron calculadas a qué sesión corresponden, dentro de las sesiones de trabajo que realiza el estudiante, con lo que el identificador del caso (Case ID), quedará etiquetado con una tupla del ID del usuario junto al número de

sesión dentro de las sesiones del usuario. En el anexo B.1 se incluye un script de Python utilizado en el cálculo de sesiones y definición de las interacciones para generar el registro de eventos, en la tabla 4.4 se muestra un ejemplo de un segmento del registro de eventos generado.

Tabla 4.4. Ejemplo de un segmento del Registro de Eventos generado para el análisis

Case ID	**Tiempo**	**Interacción**	**SRL**	**Completó**	**Sesión**
user01-1	1448567431	Video-Lecture.begin	3.162	False	1
user01-2	1448567737	Video-Lecture.complete	3.162	False	2
user01-2	1448568139	Assessment.try	3.162	False	2
user01-2	1449103918	Video-Lecture.review	3.162	False	1
user02-1	1449104348	Assessment.pass	3.433	True	1
user02-2	1449104694	Assessment.review	3.433	True	2
.....					

4.2.3. Fase 3: Descubrimiento del modelo de proceso

Para el descubrimiento del modelo de proceso, fueron utilizadas las herramientas Disco y Celonis, mencionadas en los capítulos anteriores. Sobre el modelo de proceso obtenido inicialmente con el registro de eventos completo, en la fase de análisis se identificaron las secuencias de interacción más frecuentes de los estudiantes. Con esta información, se procedió a filtrar el registro de eventos, para iterar esta fase obteniendo modelos de proceso específicos para cada tipo de secuencia de interacción más frecuente.

4.2.4. Fase 4: Análisis del modelo

Una vez generado el modelo de procesos, se analizaron e identificaron las secuencias de interacción más frecuentes de los estudiantes. Una secuencia de interacción se define como un conjunto de interacciones concatenadas (de una interacción a otra) del mismo estudiante dentro de una sesión. Es decir, el camino que un estudiante sigue a través del contenido del MOOC dentro de una sesión. Las secuencias de interacción se utilizaron en

primer lugar para un análisis exploratorio y luego para realizar agrupamiento (*clustering* en inglés).

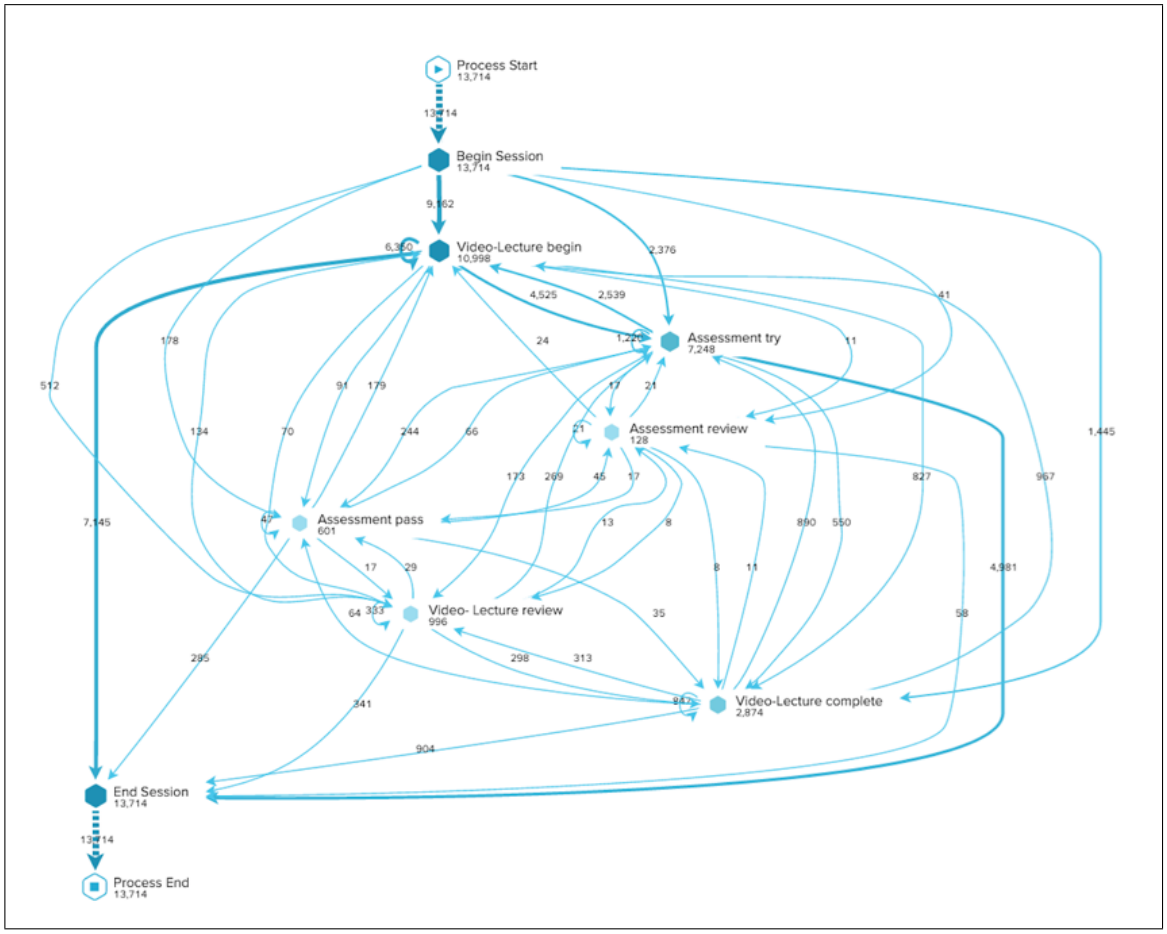


Figura 4.2. Modelo de proceso completo con todas las secuencias de interacción de los 3 MOOCs por sesión.

Como primer resultado de aplicar los algoritmos de descubrimiento, se obtuvo un modelo de procesos tipo “espagueti” (Figura 4.2). El modelo de proceso de espaguetis es un término utilizado en PM para referirse a un modelo con muchos arcos y cruces, en el cual es difícil entender u observar patrones, lo que se debe a que los datos considerados en la construcción del modelo reflejan un proceso no estructurado (Van der Aalst, 2016).

La visualización del modelo de procesos está delimitada por un punto de inicio y un punto final representados con un hexágono de color blanco con una imagen inicio (*play*

en inglés) y una imagen de parada (*stop* en inglés) en su interior, respectivamente. Las distintas interacciones están representadas con un hexágono relleno de colores. Los arcos y flechas conectan dos o más interacciones a lo que se denomina secuencias de interacción que fueron realizadas por diferentes estudiantes. Por ejemplo, una secuencia de interacción sería desde *Iniciar sesión (→) videoconferencia-comienzo a (→) Finalizar sesión*[1], lo que indica que un estudiante comenzó una sesión, luego vio una video-lectura y luego terminó una sesión. La Figura 4 muestra un subconjunto de secuencias de interacción extraídas del modelo de proceso principal para proporcionar una mejor explicación sobre su semántica. El modelo de procesos también contiene números junto a cada hexágono. Estos números indican el número de veces que la interacción indicada en el hexágono se repitió en todas las sesiones del conjunto de datos. Por ejemplo, la figura 4.2 muestra que el registro de eventos contiene 13714 interacciones de Iniciar sesión; es decir, hubo 13714 sesiones registradas en el conjunto de datos. Los números sobre los arcos con flechas indican el número de secuencias de interacción de las dos interacciones interconectadas que se han identificado dentro de una sesión, y las flechas indican la dirección. La figura 4.3 muestra que la secuencia de interacción de *Iniciar sesión a (→) Video-lecture-begin* se realizó 9162 veces. Esto significa que, de las 13714 sesiones iniciadas, 9162 secuencias de interacción se realizaron hacia *Video-lecture-begin*.

Una vez generado el modelo de procesos, a este se le aplican filtros de acuerdo a los niveles de SRL asociados a cada estudiante y a la aprobación de éstos, para así obtener los modelos que representan cada uno de estos casos, y así responder a las dos primeras preguntas de investigación planteadas. Por otra parte, se registran cuales son las distintas secuencias de actividades existentes en los datos y sus frecuencias. A cada secuencia con actividades distintas a otra se le denomina **variante**, y serán utilizadas para la clasificación de la actividad que busca responder la tercera pregunta de investigación. El detalle de cómo se realizó dicha clasificación se presenta en la siguiente sección.

[1]Es necesario aclarar que interacciones *Iniciar Sesión* y *Finalizar Sesión*, no son explícitas en los datos extraídos de Coursera, sino que son inferidos desde la actividad, esto buscando hacer más comprensible el modelo de procesos de la sesión, y permitiendo diferenciar como transiciones cuando se inicia o termina una sesión con cierto tipo de actividad.

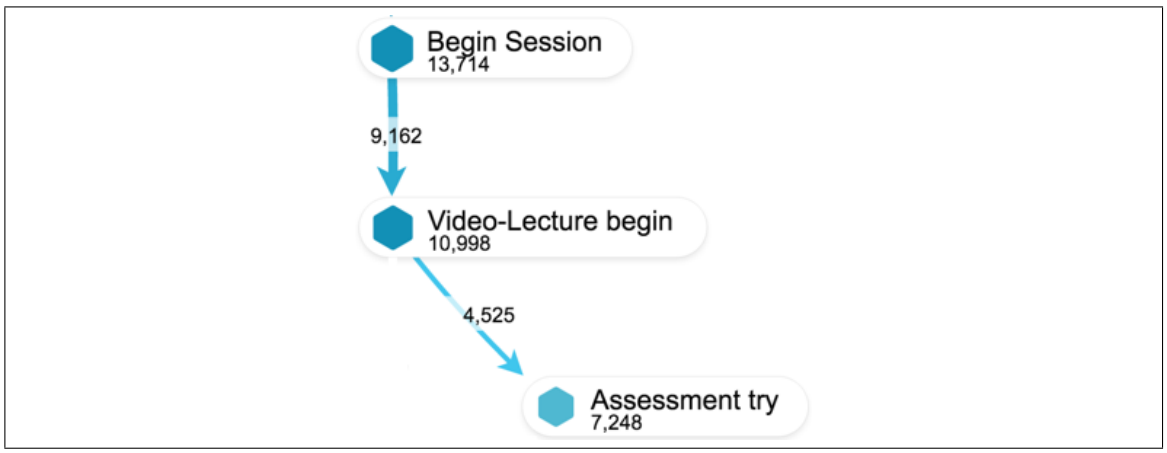


Figura 4.3. Ejemplo de representación de secuencias de interacción extraídas del modelo de procesos completo.

Una vez generado el modelo de procesos general, se aplicaron filtros al registro de eventos para así iterar sobre la fase de descubrimiento del modelo, y obtener modelos de proceso más específicos. Con esto fue posible extraer información para responder a las tres preguntas de investigación planteadas en este estudio:

RQ1. ¿Es posible utilizar minería de procesos para identificar las secuencias de interacción más frecuentes de los estudiantes que reflejen estrategias de aprendizaje específicas? En caso afirmativo, ¿qué tipos de interacciones pueden identificarse? Para responder a esta pregunta, se analizaron los modelos de procesos en la etapa de análisis del modelo para identificar los patrones de secuencias de interacción más frecuentes. Para esto, como paso inicial se analizó primero, los modelos, considerando todos los datos de los tres cursos, y después se consideró solo los datos de cada curso por separado.

RQ2. ¿En qué se diferencian las secuencias de interacción de los estudiantes con diferente rendimiento académico? Después de haber identificado los patrones de secuencia de interacción más comunes entre los participantes del MOOC en una sesión, se analizó cómo estos patrones varían según si los participantes completaron o no el curso. Para lograrlo, se filtró el archivo de registro por aquellos que completan (n = 258) y que

no completan (n = 3200). Esto nos permitió observar las diferencias entre los diferentes patrones de secuencias de interacción. Se generaron modelos de proceso para los que completan y los que no lo hacen.

RQ3. ¿En qué se diferencian las secuencias de interacción entre los estudiantes con diferentes perfiles SRL? Para responder a esta pregunta, se utilizó una técnica de clusterización jerárquico aglomerativa para agrupar a los estudiantes (N = 3458) basados en los patrones de secuenciación de interacción identificados (por ejemplo, estrategias de aprendizaje). Es decir, agrupamos a los estudiantes en grupos de acuerdo con su uso diferenciado de las estrategias de aprendizaje. Fueron utilizadas las puntuaciones obtenidas a través del cuestionario de auto-reporte de SRL, para observar cómo se distribuyen los alumnos entre los diferentes grupos.

4.3. Resultados

4.3.1. Patrones de comportamiento encontrados

A partir del modelo de proceso obtenido, la figura 4.2 muestra un total de 13714 sesiones realizadas por los estudiantes. Estas fueron clasificadas y se identificó 1956 tipos de sesiones diferentes, cada una de las cuales contenía un conjunto de secuencias de interacción que caracterizaban dicha sesión. De ahora en adelante a cada uno de estos tipos de sesión se les denominará variante. La figura 4.4 muestra una captura de pantalla del software Disco, que proporciona una lista de las 1956 variantes y una visión general de sus secuencias de interacción relacionadas a una sesión en específico, en la figura se puede observar que la variante 21 muestra 4 interacciones (eventos) con 3 secuencias de interacción y el tiempo asociado a la duración de la sesión.

Las distintas sesiones fueron ordenadas desde las más comunes hasta las menos comunes. La sesión más común fue asignada a una categoría que describe un patrón de secuencia de interacción. Por ejemplo, se analizaron los primeros tipos de sesiones más

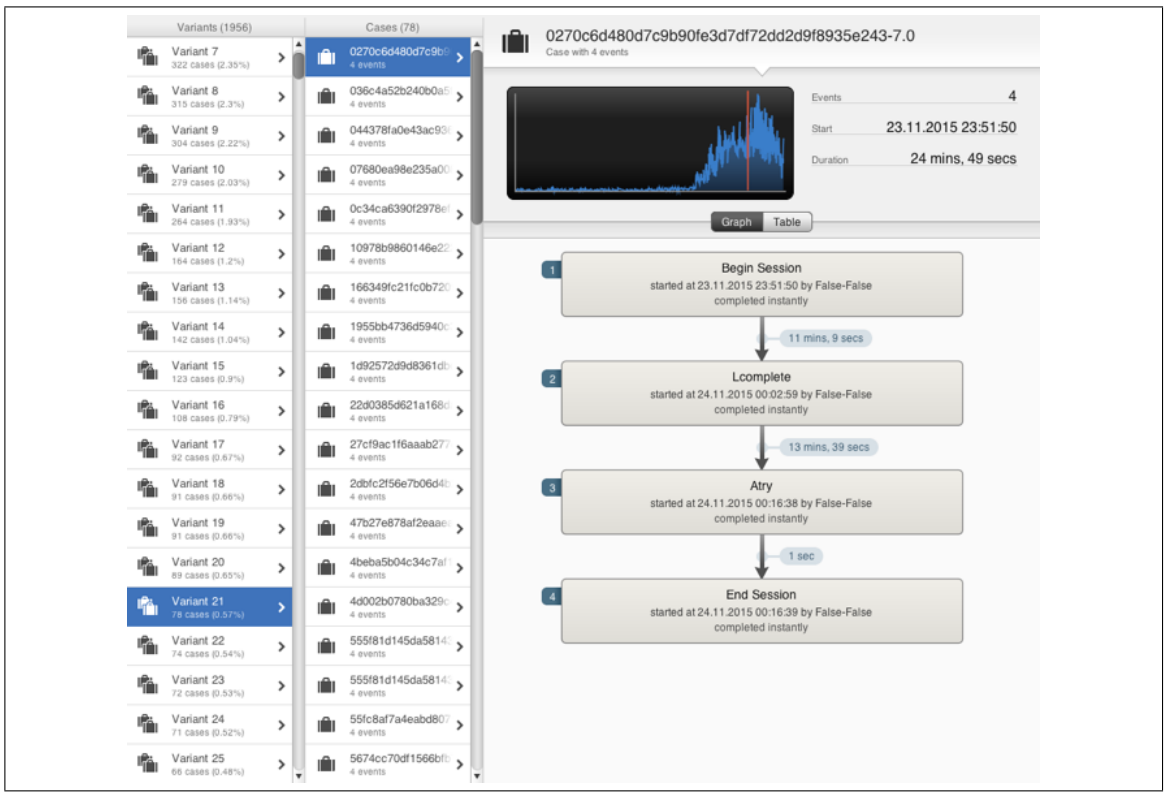


Figura 4.4. Listado de las 1956 variantes de sesión obtenidas con el software Disco.

comunes y se observó que éstas consisten en secuencias de interacción de iniciar-video-lectura (video-lecture-begin). Por lo tanto, se definió un patrón de solo-video-lecturas (Only video-lecture). A continuación, en el archivo que contiene las sesiones, se filtra en Disco y se marca este tipo de sesiones. Luego se repite el procedimiento, identificando el resto de los tipos de sesiones que quedan sin marcar en el archivo que contiene las sesiones. Este proceso fue realizado por medio de un script desarrollado en Python que une las distintas particiones resultantes al filtrar el registros de eventos, el que se incluye en el anexo B.2. Como resultado, se obtuvo los siguientes siete patrones de secuencias de interacción (se han mantenido los nombres de los patrones en inglés por nomenclatura y una mejor comprensión):

1. Only Video-lecture.
2. Only Assessment

3. Explore.
4. Assessment try to Video-lecture.
5. Video-lecture complete to Assessment try.
6. Video-lecture to Assessment complete.
7. Others.

Aquellos tipos de sesiones que encajan en múltiples patrones de secuencias de interacción (dado que son largas y dispersas) o no encajan en ningún patrón de secuencias de interacción, se clasificaron como otros. La descripción de cada patrón de interacción se basa en, si una sesión sólo contiene un cierto tipo de interacción (definida en la tabla 4.3) o si la sesión contiene un cierto tipo de secuencias de interacción entre las interacciones que son importantes en el proceso de aprendizaje (por ejemplo, intentar pasar una evaluación y luego hacer una video-lección representa cómo el estudiante busca la información faltante después de intentar pasar la evaluación). Una vez que los patrones de interacción más comunes fueron extraídos del modelo de procesos principal (figura 4.2), se definió para cada patrón un modelo de procesos (figuras 4.5, 4.6, 4.7, 4.8, 4.9, 4.10 y 4.11), con el fin de observar el comportamiento del estudiante como resultado de la interacción con el contenido de MOOC en una sesión. A continuación, se describen los siete patrones de secuencia de interacción distintos extraídos utilizando minería de procesos como sigue:

(1) ***Only Video-lecture:*** patrón de secuencia de interacción dedicado únicamente a ver video-lecturas, en las que las secuencias de interacción más comunes son *Begin session* a *video-lecture-begin* o *video-lecture-complete* o *video-lecture-review* y combinaciones de estas antes de *End session* (figura 4.5).

(2) ***Only Assessment:*** patrón de secuencia de interacción dedicado a trabajar solo con evaluaciones en las cuales la secuencia de interacción más frecuente es *Begin session* a *assessment-try* o *assessment-pass* o *assessment-review* y combinaciones de estas antes de terminar la sesión (*End session*) (figura 4.6).

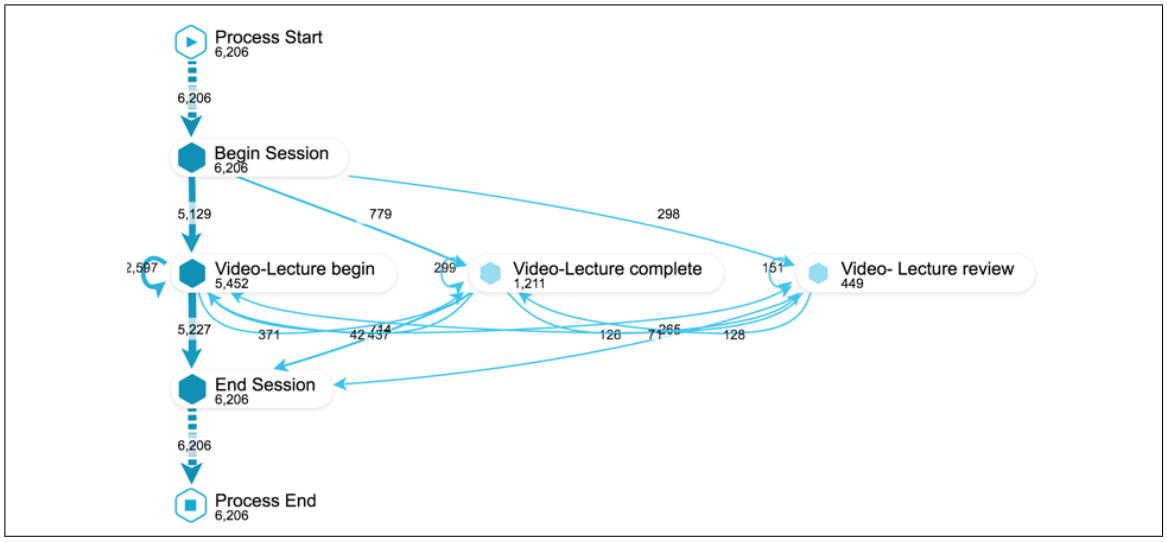


Figura 4.5. Modelo de proceso del patrón de interacción Only Video-lecture.

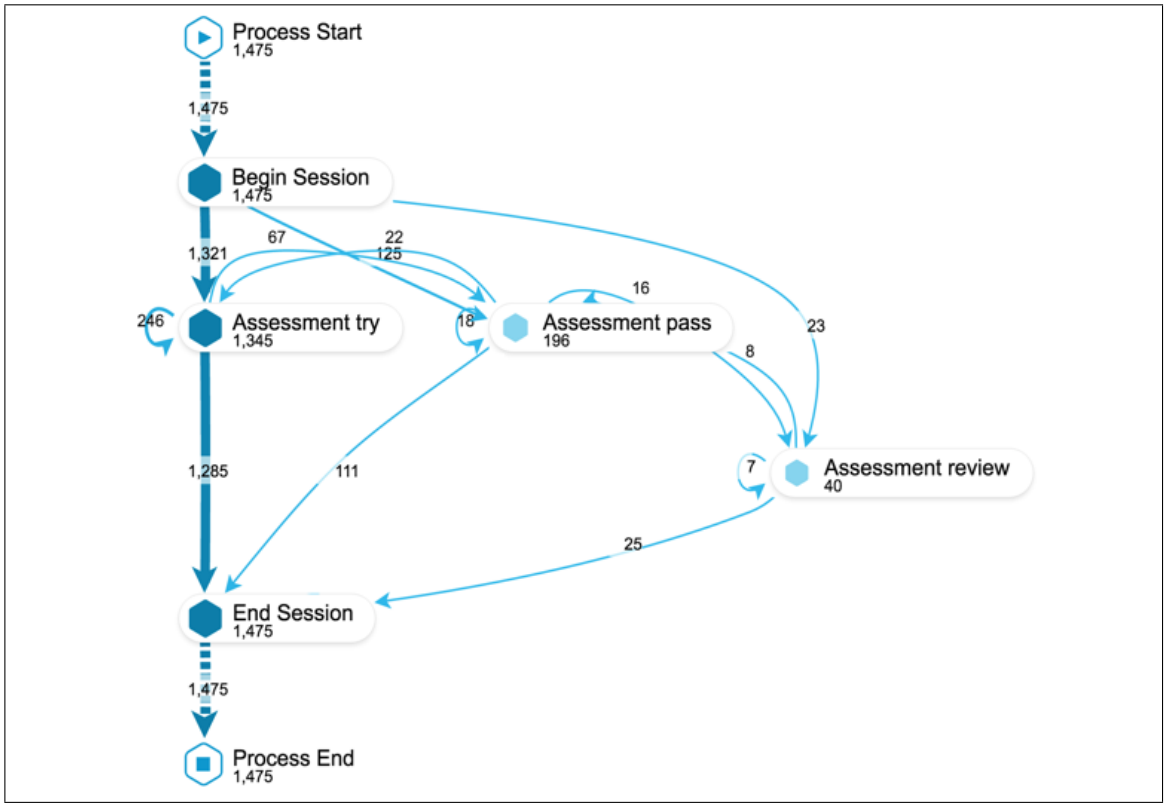


Figura 4.6. Modelo de proceso del patrón de interacción Only Assessment.

(3) ***Assessment-try to Video-lecture:*** patrón de secuencia de interacción donde las secuencias de interacción más comunes observadas son (a) *Begin session* a *Assessment-try* (con la intención de intentar resolver una evaluación) luego a *Video-lecture-begin* (buscando información en una nueva video-lectura) luego a *Assessment-try* y *End session*, (b) *Begin session* a *Assessment-try* luego a *Video-lecture-complete* (consumiendo la información de la video-lectura) luego a *Assessment-try* y *End session*, y (c) *Begin session* a *Assessment-try* luego a *Video-lecture-review* (buscando información específica) luego a *Assessment-try* y *End session* (figura 4.7).

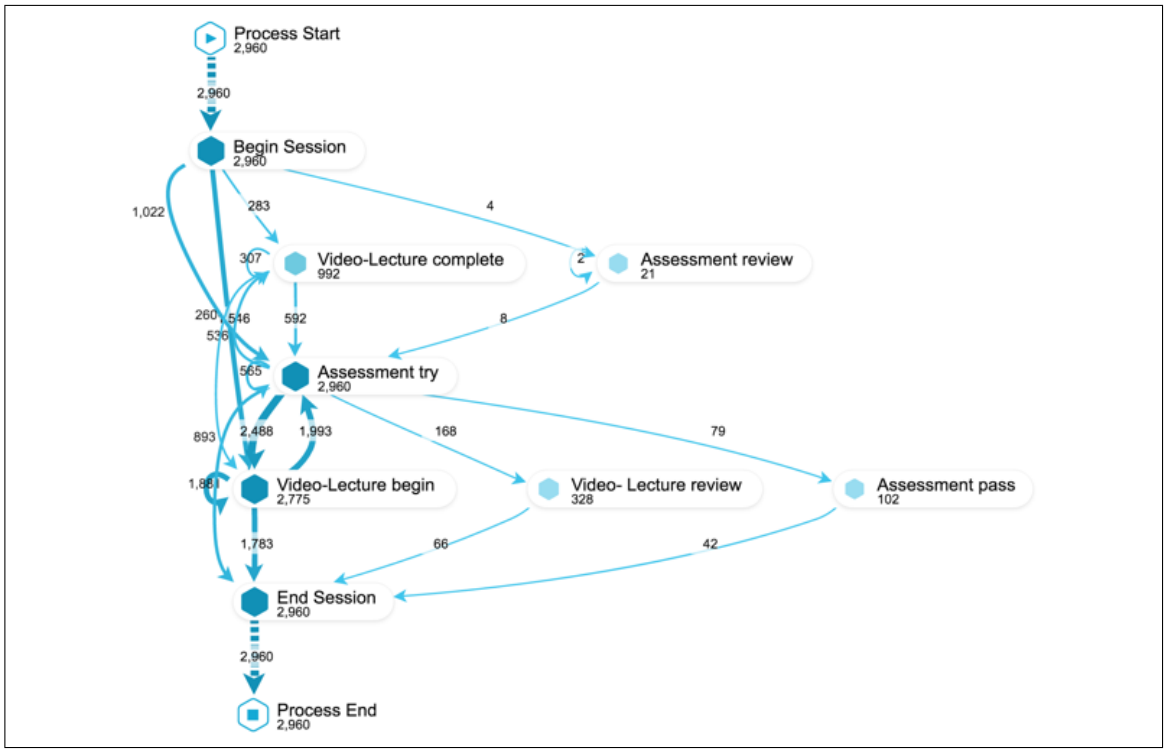


Figura 4.7. Modelo de proceso del patrón de interacción Assessment-try to Video-lecture.

(4) ***Explore:*** patrón de secuencia de interacción compuesto de un *assessment-try* y un *video-lecture-begin*, donde los estudiantes inspeccionan de forma superficial los contenidos sin intención de completarlos (figura 4.8).

(5) ***Video-lecture-complete to Assessment-try:*** patrón de secuencia de interacción donde la secuencia de interacción más común que se observa es *Begin session*

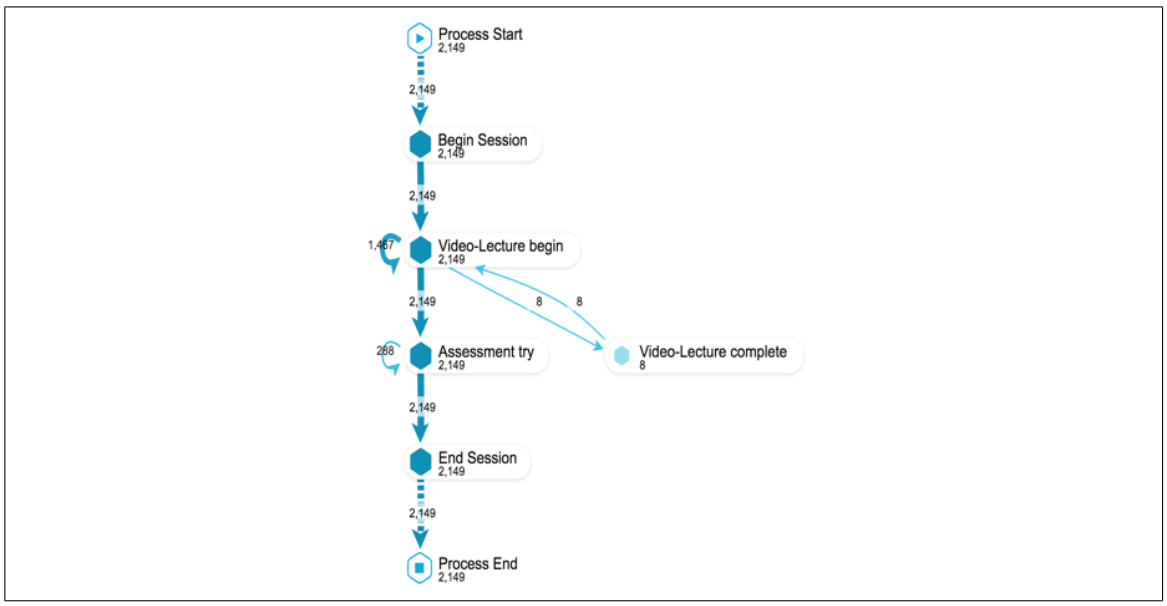


Figura 4.8. Modelo de proceso del patrón de interacción Explore.

a *Video-lecture-complete* luego a *Assessment-try* (sin lograr completarla y sin más intentos por completarla) y luego *End session* (figura 4.9).

(6) ***Video-lecture to Assessment-pass:*** patrón de secuencia de interacción donde las secuencias de interacción más comunes que se observan son (a) *Begin session* a *Video-lecture-begin* luego a *Assessment-pass* y luego *End session*, (b) *Begin session* a *Video-lecture-complete* luego a *Assessment-pass* y luego *End session*, (c) *Begin session* a *Video-lecture-review* luego a *Assessment-pass* y luego *End session*, y (d) *Begin session* a *Video-lecture-begin* luego a *Assessment-try* luego a *Assessment-pass* y luego *End session* (figura 4.10).

(7) ***Others:*** patrones de secuencia de interacción largos y dispersos que no encajan en ningún patrón de secuencia de interacción mencionado anteriormente. Las secuencias de interacción más comunes observadas son (a) *Begin session* a varias *Video-lecture-begins* luego a *Assessment-try* y luego *End session* (figura 4.11).

Los cuatro patrones más comunes de secuencias de interacción entre los estudiantes de los MOOCs (93.26 % de las sesiones registradas) son los siguientes, en orden de frecuencia: (1) ***Only Video-lecture*** (45.25 % de las sesiones siguen este tipo de patrón). La

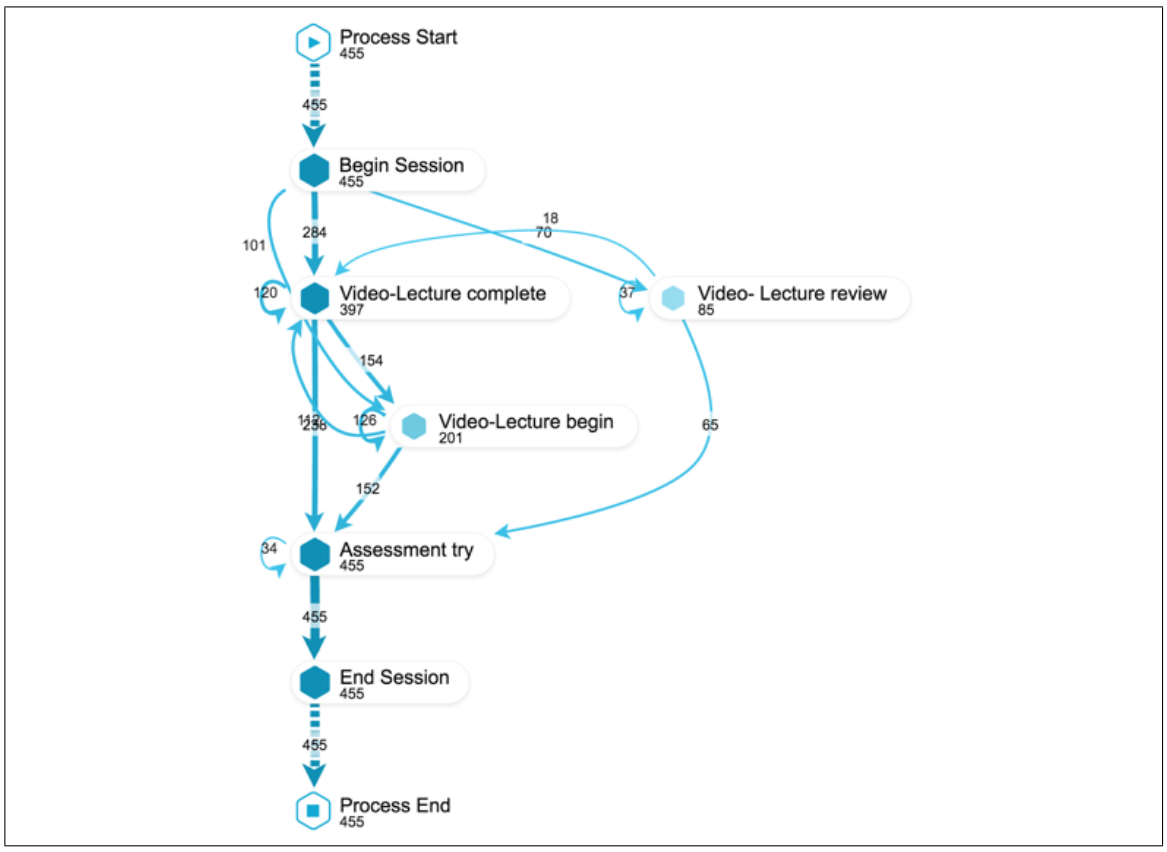


Figura 4.9. Modelo de proceso del patrón de interacción Video-lecture-complete to Assessment-try.

secuencia de interacción más común en este tipo de patrón de interacción es *Begin session*, luego *Video-lecture-begin*, luego *End session* sin completar la video-lectura; **(2) *Assessment try* → *Video-lecture*:** 21.58 % de las sesiones siguen este tipo de patrón, siendo la secuencia de interacción más común de este patrón de interacción un bucle de repetición entre *Begin session* → *Assessment-try* → *Video-lecture-begin* → *Assessment-try* → *Video-lecture-complete* → *Assessment-try* → *End session*; **(3) *Explore*:** 15.67 % de las sesiones siguen este tipo de patrón, en que el comportamiento más común de los estudiantes es seguir una secuencia de interacción desorganizada en la que van desde un tipo de contenido (evaluaciones o video-lecturas) a otros sin completarlos; **(4) *Only Assessment*:** 10.76 % de las sesiones siguen este tipo de patrón, en el que la secuencia de interacción más común

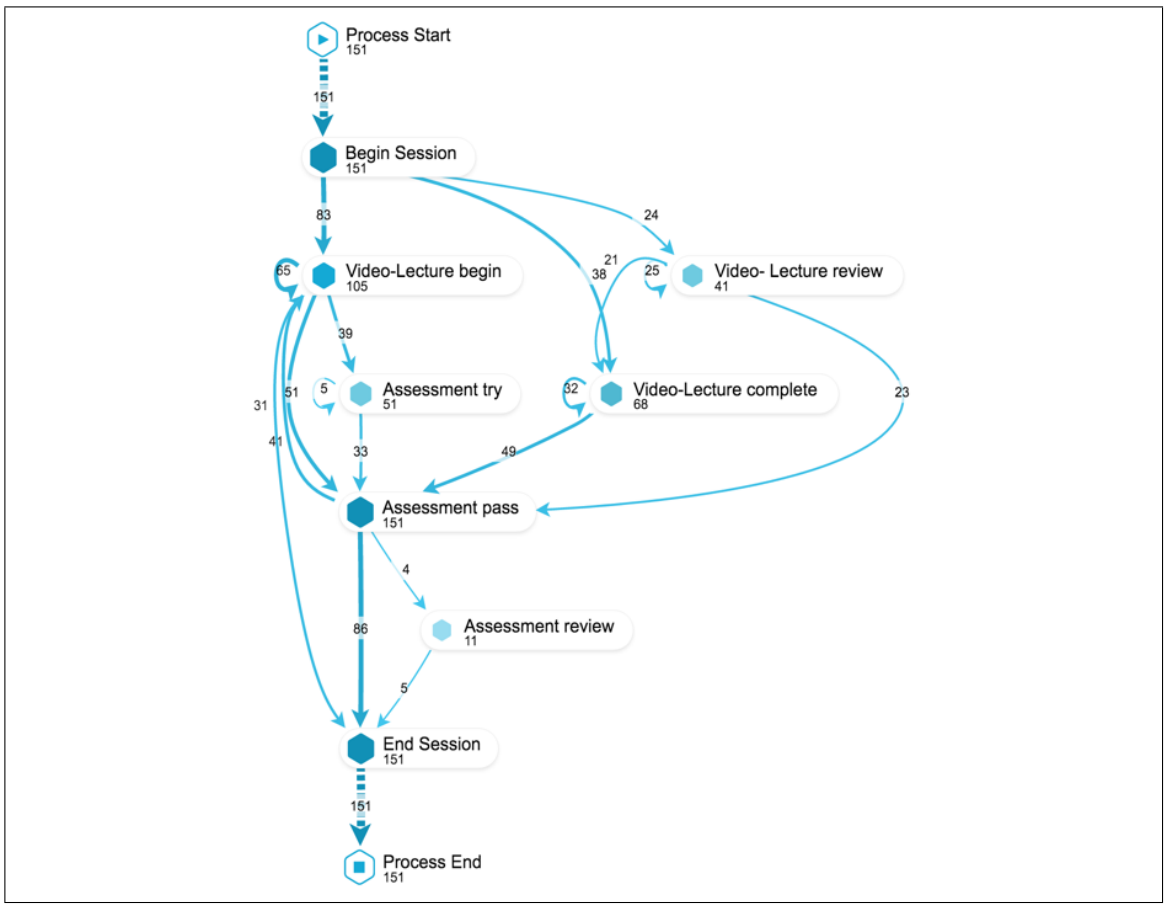


Figura 4.10. Modelo de proceso del patrón de interacción Video-lecture to Assessment-pass.

es *Begin session* → *Assessment-try* → *End-session* sin completar la evaluación. Finalmente, los patrones ***Video-lecture complete*** → ***Assessment-try*** (3.32 %), ***Video-lecture*** → ***Assessment-pass*** (1.10 %) y ***Others*** (2.32 %) son los menos comunes. Estos patrones nos ayudan a entender cómo se comportan los estudiantes en una sesión de estudio, ya sea que completen el curso o no.

En la tabla 4.5 se muestran la cantidad de sesiones que clasifican en cada patrón de interacción, y el número de estudiantes que exhibe comportamiento dentro de ese patrón de interacción. Cabe notar que un estudiante puede tener sesiones que caen dentro de patrones de interacción distintos a lo largo del curso.

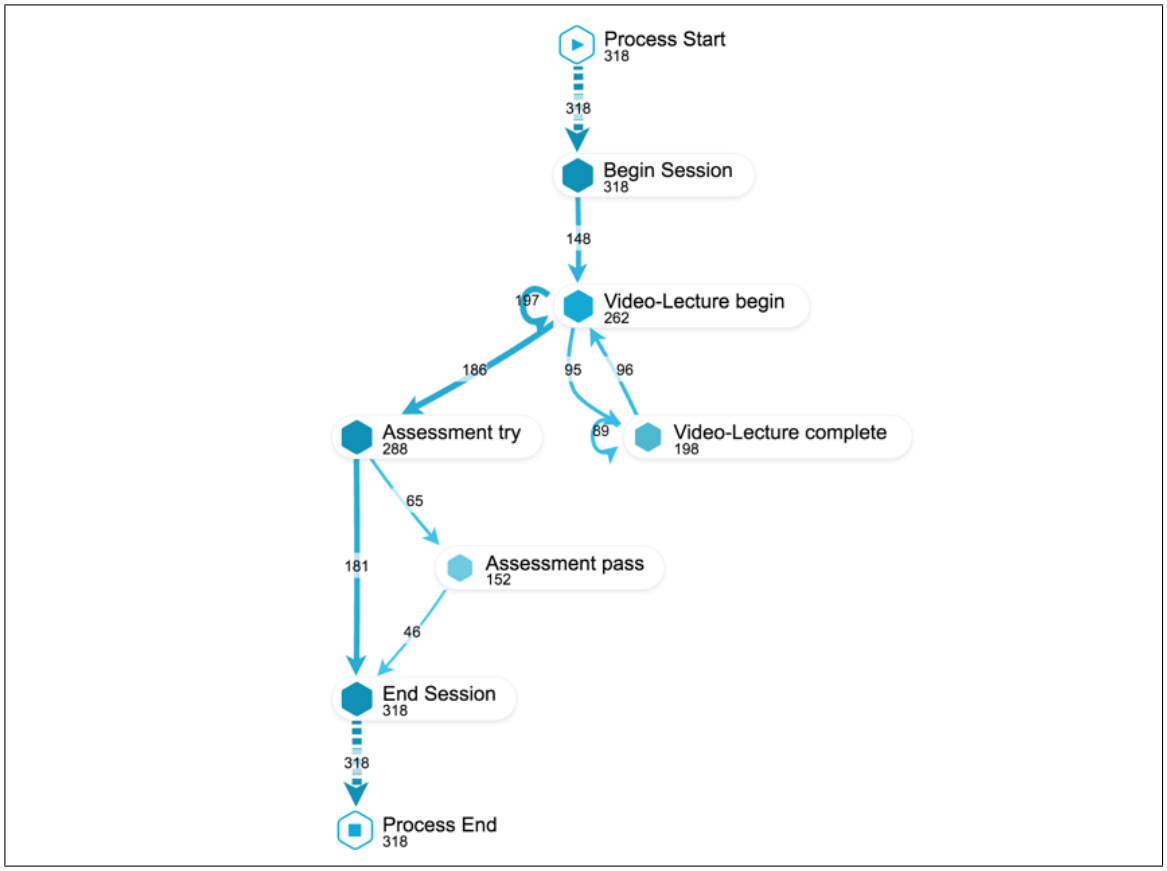


Figura 4.11. Modelo de proceso de otras sesiones, sin patrón definido.

Tabla 4.5. Proporciones de los patrones de secuencia de interacción basados en el número de sesiones (N_sesiones =13.714) realizado por los estudiantes en los 3 MOOCs y derivado de los modelos de procesos.

Patrones de secuencia de interacción	**3 MOOCS**	
	N_sesiones	**Estudiantes**
Only Video-lecture	6206 (45.25 %)	2495 (72.15 %)
Assessment try → Video-lecture	2960 (21.58 %)	1271 (36.76 %)
Explore	2149 (15.67 %)	1195 (34.56 %)
Only Assessment	1475 (10.76 %)	865 (25.01 %)
Video-lecture complete → Assessment try	455 (3.32 %)	358 (10.35 %)
Video-lecture → Assessment pass	151 (1.10 %)	132 (3.82 %)
Others	318 (2.32 %)	258 (7.46 %)
Total	**13714 (100 %)**	**-**

4.3.2. Relación entre niveles de aprobación y patrones de comportamiento observados en estudiantes.

Después de haber identificado los patrones de secuencia de interacción más comunes entre los estudiantes del MOOC en una sesión, se analizó cómo estos patrones varían según si el grupo de estudiantes completa o no el curso. Específicamente, se buscaron diferencias en los patrones de secuencia de interacción que realizaron los estudiantes que completaron, lo que debería ayudar a revelar cómo su comportamiento impacta en su aprendizaje y cómo este se relaciona con las estrategias de SRL.

Se analizaron los patrones de secuencia de interacción por sesión. **Se encontró que los estudiantes que completan el curso realizaron sesiones que contenían más evaluaciones en comparación con los estudiantes que no completan el curso.** Las sesiones de los estudiantes que completan el curso principalmente se componen de (a) intentar resolver una evaluación después de otra (llamado *Only Assessment*) o (b) intentar una evaluación y luego ver una video-lectura (llamado *Assessment try* $\rightarrow$ *Video-lecture*) o (c) mirar una video-lectura o intentar una evaluación sin completar ninguna de las dos (llamada *Explore*).

En contraste las sesiones de los estudiantes que no completan el curso principalmente se componen de mirar una video-lectura después de otra (llamado *Only Video-lecture*). Se encontraron diferencias estadísticas significativas entre el porcentaje de sesiones de cada tipo realizadas por estos dos tipos de grupos de estudiantes (tabla 4.6).

Tabla 4.6. Proporciones de los patrones de secuencia de interacción basados en el número de sesiones (N_sesiones =13714) realizado por los estudiantes en los 3 MOOCs y derivado de los modelos de procesos para los que completan el curso y los que no.

Patrones de secuencia de interacción	**Completers**		**Non-Completers**		χ^2	**valor-p**	**r**
	N_sessions	**%**	**N_sessions**	**%**			
Only Video-lecture	1253	**36.29**	4953	**48.27**	149.26	<**0.001***	0.1043
Assessment try → Video-lecture	922	**26.70**	2038	**19.86**	71.42	<**0.001***	0.0722
Explore	610	**17.67**	1539	**15.00**	13.94	<**0.001***	0.0319
Only Assessment	417	**12.08**	1058	**10.31**	8.43	<**0.01***	0.0248
Video-lecture complete → Assessment try	111	**3.21**	344	**3.35**	0.16	0.690	0.0034
Video-lecture → Assessment pass	44	**1.27**	107	**1.04**	1.26	0.262	0.0096
Others	96	**2.78**	222	**2.16**	4.34	**0.036***	0.0178
Total	**3453**	**100 %**	**10261**	**100 %**	**-**	**-**	

4.3.3. Perfiles de estudiante encontrados según actividad en cursos MOOC y su relación con los niveles de SRL auto-reportados

Para responder a esta pregunta, se empezó por clusterizar a los estudiantes (N = 3.458) en base a los patrones de secuencia de interacción identificados. Para esto se utilizó un *clustering* jerárquico aglomerativo, basado en el método de Ward. Esta técnica de clusterización ha sido utilizada anteriormente para detectar grupos de estudiantes en contextos en línea (Kovanović et al., 2015), las dimensiones consideradas para la clusterización fueron la cantidad de veces que el estudiante realizó sesiones por cada patrón de interacción. Para seleccionar el número óptimo de grupos o clústeres se inspeccionó el dendograma resultante y las diferentes formas de cortar la estructura del árbol fueron comprobadas, con el fin de obtener un número mínimo de clústeres interpretables que expliquen el comportamiento del usuario (Jovanović et al., 2017). Se probaron también otras técnicas de *clustering* como la mezcla gaussiana (*Gaussian mixture* en inglés) y *K-means* para definir el número apropiado de clústeres basados en la puntuación de silueta. Esto permitió seleccionar la solución con 3 clústeres como mejor solución (figuras 4.12 y 4.13), dado que aumentar el número de clústeres no incrementaba la puntuación de silueta.

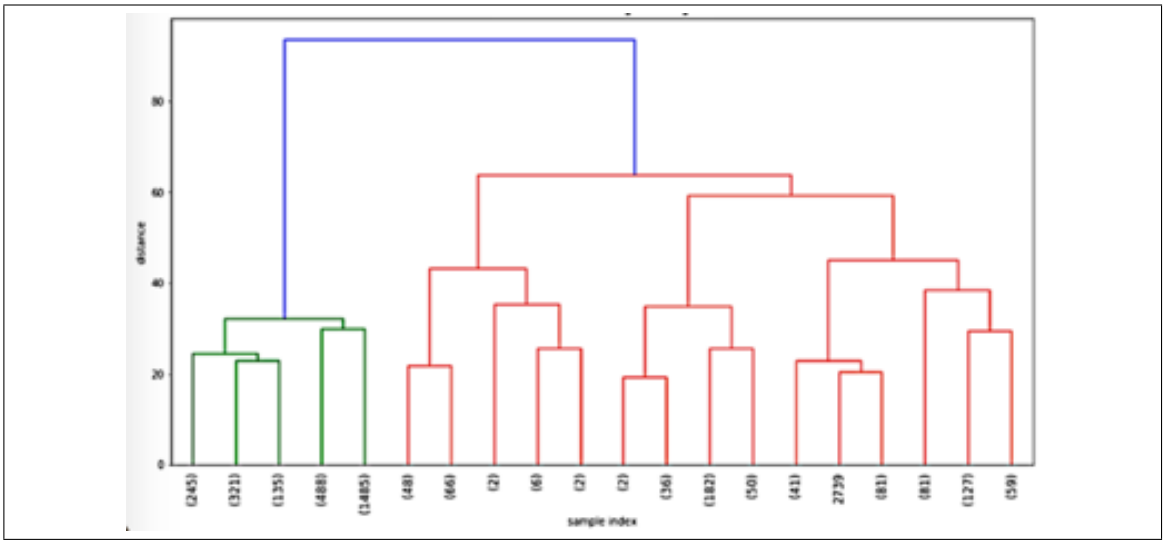


Figura 4.12. Dendograma obtenido mediante agrupación jerárquica aglomerativa.

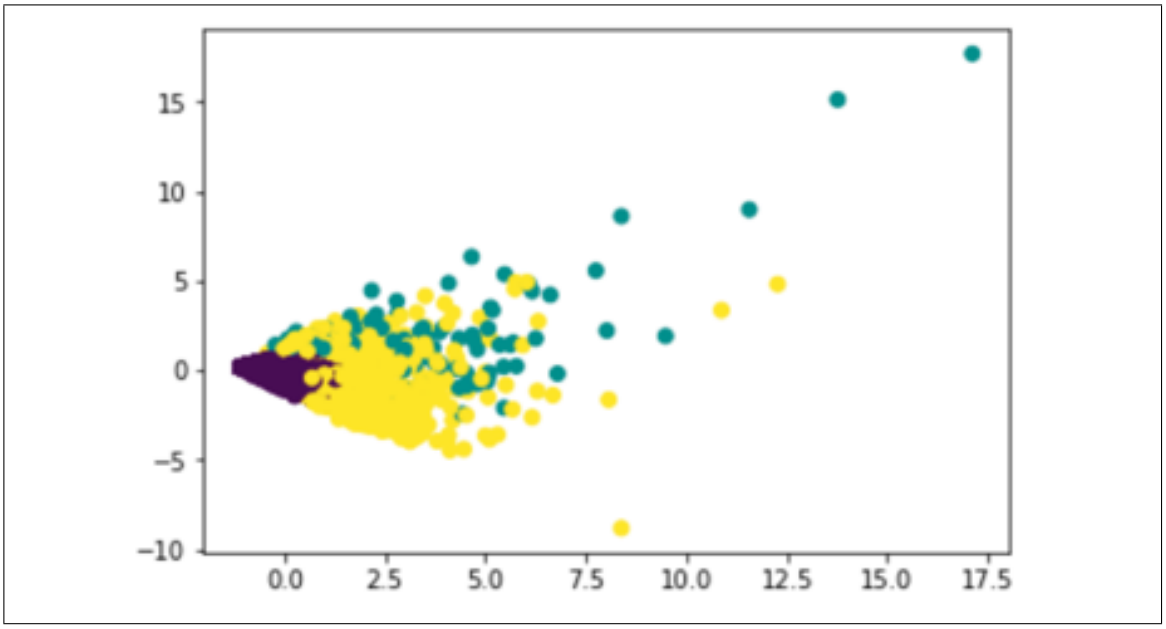


Figura 4.13. Diagrama de dispersión con puntuación de la silueta = 0,5320, para una reducción de dimensionalidad a 2 dimensiones.

Como resultado, la tabla 4.7 describe los clústeres resultantes en términos de (1) los seis patrones de secuencia de interacción identificados (se ha descartado el patrón de secuencia de interacción otros como variable) usados para el clustering; (2) la puntuación de SRL obtenida del cuestionario; y (3) la finalización del curso.

A partir de la tabla 4.7 se ha analizado la similitud en los perfiles de SRL entre cada grupo de clústeres. Como resultado, no se observaron diferencias estadísticamente significativas entre el Clúster 2 y 3, mientras que se observaron diferencias estadísticamente significativas al comparar con el Clúster 1. La tabla 4.8 y la tabla 4.9 muestran las diferencias entre cada clúster basadas en la puntuación del perfil de SRL, en la tabla 4.8 se utiliza un nivel de confianza del 95 %.

Los grupos resultantes que emergen de los datos, indican diferentes tipos de estrategias de aprendizaje que los estudiantes han adoptado mientras se enfrentan al MOOC. Si se busca por diferencias específicas entre los diferentes patrones de secuencia de interacción realizados por cada clúster, es posible describirlos como sigue:

Tabla 4.7. Estadísticas de resumen para los tres grupos (clústeres) de estudiantes: mediana y desviación estándar entre paréntesis.

Patrones de secuencia de interacción	Clúster 1	Clúster 2	Clúster 3
Only Video-lecture	4.67 (5.41)	22.57 (33.79)	15.72 (13.13)
Assessment try → Video-lecture	3.39 (7.09)	19.85 (18.60)	19.52 (21.42)
Explore	1.84 (3.61)	8.61 (9.40)	10.18 (11.37)
Only Assessment	0.65 (1.62)	4.18 (5.39)	4.39 (6.04)
Video-lecture complete → Assessment try	0.00 (0.00)	1.75 (3.70)	3.84 (4.95)
Video-lecture →Assessment pass	0.00 (0.00)	8.70 (6.05)	0.09 (0.80)
Puntaje de SRL	3.06 (0.51)	3.12 (0.49)	3.11 (0.52)
Estudiantes	2674 (77.32 %)	124 (3.59 %)	660 (19.09 %)
Completan	22 (0.8 %)	36 (29.03 %)	200 (30.30 %)
No Completan	2652 (99.2 %)	88 (70.97 %)	460 (69.70 %)

Tabla 4.8. Diferencias entre cada clúster basadas en la puntuación del perfil SRL

Clúster	Clúster	t	valor-p
2	3	0.1030	0.9179
1	2-3	-27.333	0.0063*

Clúster 1 – Estudiantes exploradores: este grupo está compuesto por estudiantes con puntuaciones de SRL más bajas en comparación con los otros dos grupos. Los estudiantes de este clúster en promedio por sesión realizan un número bajo de video-lecturas y en promedio por sesión realizan pocos intentos para tratar de resolver las evaluaciones. Estos estudiantes tienen una baja actividad en el curso (generalmente los estudiantes de este grupo ven una sola video-lectura o comienzan a explorar al principio del curso investigando sobre los contenidos de los materiales con el curso ya iniciado).

Clúster 2 – Estudiantes exhaustivos: este grupo está compuesto por estudiantes con puntuaciones de SRL superiores a las del clúster 1, por lo que pueden considerarse como más autorregulados (ver tabla 4.7). Los estudiantes en este grupo despliegan una variedad de estrategias de aprendizaje por sesión. Además, miran más video-lecturas en promedio

por sesión que los estudiantes de los otros clústeres. A partir de las secuencias de interacción observadas, los estudiantes de este grupo tienden a seguir la ruta de aprendizaje que proporciona la estructura del curso. También invierten más tiempo mirando video-lecturas y, por lo tanto, muestran un mayor nivel de compromiso que los estudiantes del clúster 3. Por lo tanto, los estudiantes de este grupo se concentran en realizar patrones de secuenciación de interacción en un orden específico que les permite aprender en profundidad el contenido del curso.

Clúster 3 – Estudiantes estratégicos: este grupo está compuesto por estudiantes con puntuaciones de SRL similares a las del clúster 2, lo que sugiere que la diferencia en el comportamiento observado no se debe a diferencias en sus perfiles SRL, pero sí en el uso de distintas estrategias de aprendizaje. Los estudiantes del clúster 2 y clúster 3 tienen tasas de finalización similares (29 % y 30 % respectivamente). Sin embargo, los estudiantes del clúster 3 ven menos video-lecturas y completan más evaluaciones en promedio por sesión. Además, los estudiantes del clúster 3 tienden a explorar más los contenidos del curso que los estudiantes en los clústeres 1 y 2. Estas diferencias nos conducen a describir a este grupo de estudiantes como más estratégicos o más orientados a objetivos específicos. Según Biggs (2012), los estudiantes estratégicos tienden a concentrar sus esfuerzos en las evaluaciones para pasarlas (objetivos orientados al desempeño) y que les permitan aprobar el curso, demostrando de esta forma menos compromiso en general. Esta interpretación es coherente con la observación de que los estudiantes en el clúster 3 exhiben un nivel de compromiso que es inferior al del clúster 2. En la tabla 4.9 se comparan los clústeres con respecto a los patrones de secuencia de interacción, y se proporciona la información necesaria para responder a la RQ3. La tabla muestra diferencias estadísticamente significativas con un nivel de significación de 0,05 (**) para los patrones *Only Video-lecture*, *Video-lecture complete* $\rightarrow$ *Assessment try* y *Video-lecture* $\rightarrow$ *Assessment pass*; y diferencias estadísticamente significativas con un nivel de significación de 0,1 (*) para el patrón *explore*, con tamaños de efectos (r) que van desde pequeños (*Only Video-lecture*, *Explore*); medio (*Video-lecture complete* $\rightarrow$ *Assessment try*) y grande (*Video-lecture* $\rightarrow$ *Assessment pass*).

Tabla 4.9. Comparaciones entre los clúster 2 y 3 en relación a la media de patrones de secuencia de interacción realizados

Patrones de secuencia de interacción	Clúster		t	p	r
	2	3			
Only Video-lecture	22.57	15.72	22.276	0.0276**	0.1917
Assessment try → Video-lecture	19.85	19.52	0.1788	0.8583	0.0129
Explore	8.61	10.18	16.393	0.100*	0.1159
Only Assessment	4.18	4.39	0.3880	0.6984	0.0284
Video-lecture complete → Assessment try	1.75	3.84	54.396	<0.001**	0.3476
Video-lecture → Assessment pass	8.70	0.09	158.244	<0.001**	0.6859

4.4. Conclusiones del caso de estudio

Se ha presentado un caso de estudio donde se aplican exitosamente las fases de la propuesta metodológica de PM para análisis de datos educacionales en cursos MOOC, si bien, dado el enfoque exploratorio del proyecto de investigación y los datos disponibles, no fue posible aplicar la totalidad de las actividades descritas en la propuesta metodológica, si se obtuvieron distintos resultados que amplían el entendimiento de cómo trabajan los estudiantes en un contexto MOOC. Las contribuciones principales de esta investigación, que fueron posibles por medio de las técnicas de PM son:

1. Identificación de seis patrones de secuencia de interacción más frecuentes que los participantes exhiben en un MOOC;
2. Diferenciación de los patrones de secuencia de interacción entre los estudiantes con un rendimiento diferente en el curso: los estudiantes que completan el curso interactuaron con mayor frecuencia con las evaluaciones en comparación con aquellos que no completan el curso;
3. Identificación de tres perfiles de estudiantes basados en sus patrones de secuencias de interacción. Estos perfiles observados, se confirman gracias a investigaciones previas sobre el comportamiento de los grupos de estudiantes: estudiantes exhaustivos que siguen la estructura secuencial esperada del MOOC; estudiantes

estratégicos que buscan la información necesaria para aprobar las evaluaciones, y estudiantes exploradores que se comportan de manera irregular y no estructurada.

Estas contribuciones responden satisfactoriamente las preguntas de investigación planteadas, lo que comprueba que las fases en que se estructura la propuesta metodológica, y las actividades ejecutadas como parte de esta, son adecuadas para la aplicación de técnicas de PM a un estudio exploratorio de actividad en MOOCs.

Además, las conclusiones nos abren distintas posibilidades de trabajo futuro. Entre estas líneas se encuentran:

- Ampliar el estudio de los patrones de secuencias de interacción ampliando el conjunto de recursos del MOOC con el que puede interactuar el estudiante como son foros, actividades de lectura, evaluaciones de tipo formativo, etc., así como también analizar por medio de un *plug-in* externo la interacción de los estudiantes con otro tipo de recursos ajenos al MOOC.
- Desarrollar modelos predictivos basados en los patrones de secuencias de interacción que permitan anticipar los resultados académicos de los estudiantes.
- La recomendación de rutas o itinerarios de aprendizaje a los estudiantes a partir de la finalización de la primera semana de estudio en un MOOC, utilizando a priori información recopilada por cuestionarios de auto-reporte y los patrones de secuencias de interacción observados.
- El uso de enfoques distintos para definir el caso o instancia del proceso, como, por ejemplo, el análisis del proceso realizado por un estudiante en el ciclo de actividad entre una evaluación completada y la siguiente evaluación completada, y cómo estos ciclos de actividad aportan información de la trayectoria del estudiante a lo largo de todo el curso.

4.5. Discusión sobre aplicación de la metodología

Tras la aplicación de la propuesta metodológica al caso de estudios presentado, se define si los resultados obtenidos resuelven las preguntas de investigación planteadas:

- RQ1. ¿Es posible adaptar y proponer una metodología basada en técnicas de PM para estudiar el comportamiento de los estudiantes a partir de datos de una plataforma MOOC?
- RQ2. ¿Qué variaciones debo realizar en la metodología de PM para extraer patrones de comportamiento y perfiles de estudiante a partir de datos de varios MOOCs?

En relación a RQ1, se concluye que sí fue posible adaptar y proponer una metodología basada en técnicas de PM para el estudiar el comportamiento de los estudiantes en los MOOCs presentados, esto debido a que la aplicación de las fases de la metodología en este caso de estudio permitió por una parte sistematizar las actividades en distintos momentos del curso como actividades comparables entre sí, y entre distintos estudiantes, y además permitió definir una clasificación de patrones de interacción, que sin ser exhaustiva, posibilita el mapeo de estos patrones con estrategias de SRL que es presentado en el artículo original de Maldonado-Mahauad et al. (2018), es decir, posibilita la asociación de los datos de comportamiento con teorías educativas que buscan explicar y describir dicha actividad.

Por otra parte, en relación a RQ2, se observa que las particularidades del contexto MOOC son principalmente relevantes para las fases 2 y 4 de la propuesta metodológica, es decir, para la creación del registro de eventos, y para el análisis del modelo de proceso obtenido.

Para la creación del registro de eventos se requieren de definiciones concretas de la instancia del proceso, que en este caso de estudio es definido como una sesión de trabajo continuo inferida desde la temporalidad de los datos, aunque la instancia del proceso podría definirse de otras formas, como en el caso de (Mukala, Buijs, Leemans, y van der

Aalst, 2015) donde la instancia corresponde al recorrido de un estudiante por el curso completo. También se requieren de definiciones en los eventos del proceso que permitan el análisis de la actividad más allá de la conformidad con cierta ruta esperada en el curso. Esto puede lograrse al definir actividades en categorías más generales, y considerarlas indistintas entre sí para efectos de explicar el comportamiento exhibido por el estudiante, que es lo realizado en este caso de estudios al modelar el proceso en base a actividades categorizadas como *video-lectures* y *assessments*, y no como actividades específicas y únicas a lo largo del curso.

En relación a la fase de análisis, el obtener conclusiones sobre el comportamiento de los estudiantes debe ser guiado ya sea implícita o explícitamente por conocimiento adicional sobre el escenario de aprendizaje estudiado y/o por teorías educativas apropiadas para explicar las conductas que se buscan estudiar, esto para identificar si la evaluación realizada sobre el modelo de procesos y las iteraciones a modelos de proceso específicos de cierto patrón de interacción efectivamente permiten resolver las preguntas de investigación. En el caso de estudio presentado los patrones de interacción además de aparecer desde los datos dada su frecuencia de ocurrencia, también tienen una interpretación teórica en términos de estrategias de SRL, que es lo que permite darle sentido posteriormente a los clústeres de estudiantes encontrados.

Finalmente se considera que, con los resultados y conclusiones obtenidos por el caso de estudio, son apropiadas para el estudio de datos educacionales de MOOCs las distintas fases de la propuesta metodológica. Si bien dadas las limitaciones particulares de los datos disponibles y de la no existencia de modelos de proceso previos al realizarse la investigación del caso de estudio, cuyo carácter fue exploratorio, no permitió la ejecución exhaustiva de todas las actividades posibles descritas en la propuesta metodológica (como por ejemplo la verificación de conformidad), queda abierta la posibilidad de aplicar dichas actividades en trabajos futuros, en la medida que el enfoque de estos trabajos así lo requiera, dado el carácter flexible de la metodología PM^2, y de la propuesta metodológica acá presentada. Con esto se considera satisfactoria la validación de la propuesta metodológica.

5. CONCLUSIONES Y TRABAJO FUTURO

5.1. Conclusiones

La incorporación de técnicas de PM al estudio de la actividad en MOOCs durante los últimos años ha sido poco estudiada y es un tema que aún se mantiene vigente, abriendo posibilidades a múltiples líneas de investigación en el área de EPM. Dentro de este contexto, la presente tesis hace una contribución clara en esta dirección por medio de 2 resultados obtenidos:

1. Una propuesta metodológica de 4 fases para extraer y analizar patrones de comportamiento en MOOCs, en la que se indican las actividades necesarias y opcionales a realizar en cada fase, y las consideraciones necesarias para el contexto MOOC. Propuesta que ha sido probada en su aplicabilidad con datos de la plataforma Coursera y es extrapolable a otras plataformas MOOC.
2. Un caso de estudio en donde es aplicada la propuesta metodológica a modo de ejemplo y validación de la aplicabilidad y uso de la herramienta. Concretamente, se presenta un caso de estudio que tiene como objetivo extraer patrones de comportamiento de autorregulación de los estudiantes de un MOOC. De la aplicación de la metodología se obtienen los siguientes patrones: a) *Only Video-lecture*, b) *Only Assessment*, c) *Explore*, d) *Assessment try* $\rightarrow$ *Video-lecture*, e) *Video-lecture complete* $\rightarrow$ *Assessment try* y f) *Video-lecture* $\rightarrow$ *Assessment pass*. Por medio del análisis de estos patrones se obtuvo que:
   a) Los patrones de secuencia de interacción entre los estudiantes que completan el curso muestran que estos interactuaron con mayor frecuencia con las evaluaciones en comparación con aquellos que no completan el curso.
   b) Se identificaron tres perfiles de estudiantes basados en sus patrones de secuencias de interacción: 1) estudiantes exhaustivos que siguen la estructura secuencial “esperada” del MOOC; 2) estudiantes estratégicos que buscan la información necesaria para aprobar las evaluaciones, y 3) estudiantes exploradores que

se comportan de manera irregular y no estructurada. Estos perfiles observados, se confirman gracias a investigaciones previas sobre el comportamiento de los grupos de estudiantes en un entorno en línea.

En relación a las preguntas de investigación planteadas para este trabajo, se concluye que sí fue posible adaptar y proponer una metodología basada en técnicas de PM para el estudiar el comportamiento de los estudiantes en cursos MOOC, validando la propuesta satisfactoriamente con el caso de estudio presentado. Dentro de las adaptaciones realizadas para la aplicación de esta propuesta metodológica para el contexto MOOC, destacan las definiciones que requiere la fase de generación del registro de eventos y la fase de análisis, al depender fuertemente de conocimiento sobre el contexto de los cursos y la plataforma, y de los modelos teóricos educativos con los cuales se busca explicar el comportamiento observado.

Estos resultados proveen de una sistematización de herramientas existentes y nuevas perspectivas sobre cómo abordar el estudio de la actividad de estudiantes en cursos MOOC. Además, se ha contribuido con nuevo conocimiento sobre cómo se desarrollan los procesos de autorregulación en entornos de aprendizaje en línea, específicamente sobre un entorno de aprendizaje MOOC. También estos resultados ayudarán a ganar un mejor entendimiento sobre cómo se deberían diseñar los MOOC y de cómo las plataformas pueden ayudar a reducir el número de estudiantes que abandonan el curso sin lograr sus objetivos.

5.2. Trabajo futuro

Como líneas de trabajo futuras relativas a la aplicación de PM para estudio de actividad en MOOCs, se tiene considerado:

- Validación de la metodología por terceros, es decir, que otros investigadores apliquen la metodología en otros contextos con otros MOOCs para entender mejor su aplicabilidad a otros contextos.

- Validación con expertos de PM, para ver qué piensan y qué cambiarían de la propuesta metodológica.
- Utilización de la propuesta metodológica en diversos enfoques de investigación aplicables al contexto MOOC, como es indicado en las conclusiones del caso de estudio.

BIBLIOGRAFÍA

Aggarwal, C. C., y Wang, H. (2011). Text mining in social networks. En *Social network data analytics* (pp. 353–378). Springer.

Agrawal, R., y Srikant, R. (1995). Mining sequential patterns. En *Data engineering, 1995. proceedings of the eleventh international conference on* (pp. 3–14).

Bannert, M., Reimann, P., y Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and learning*, *9*(2), 161–185.

Barnard, L., Paton, V., y Lan, W. (2008). Online self-regulatory learning behaviors as a mediator in the relationship between online course perceptions with achievement. *The International Review of Research in Open and Distributed Learning*, *9*(2).

Barreiros, B. V., Lama, M., Mucientes, M., y Vidal, J. C. (2014). Softlearn: A process mining platform for the discovery of learning paths. En *Advanced learning technologies (icalt), 2014 ieee 14th international conference on* (pp. 373–375).

Beheshitha, S. S., Gašević, D., y Hatala, M. (2015). A process mining approach to linking the study of aptitude and event facets of self-regulated learning. En *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 265–269).

Biggs, J. (2012). What the student does: teaching for enhanced learning. *Higher Education Research & Development*, *31*(1), 39–55.

Bogarín, A., Cerezo, R., y Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(1).

Bogarín, A., Romero, C., Cerezo, R., y Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. En *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 11–15).

Bose, R. J. C., y Van der Aalst, W. M. (2009). Abstractions in process mining: A taxonomy of patterns. En *International conference on business process management* (pp. 159–175).

Bozkaya, M., Gabriels, J., y van der Werf, J. M. (2009). Process diagnostics: a method based on process mining. En *Information, process, and knowledge management, 2009. eknow'09. international conference on* (pp. 22–27).

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., y Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edx's first mooc. *Research & Practice in Assessment*, *8*.

Cairns, A. H., Gueni, B., Fhima, M., Cairns, A., David, S., y Khelfa, N. (2015). Process mining in the education domain. *International Journal on Advances in Intelligent Systems*, *8*(1).

Cooper, S., y Sahami, M. (2013). Reflections on stanford's moocs. *Communications of the ACM*, *56*(2), 28–30.

Cormier, D. (2008). The cck08 mooc–connectivism course, 1/4 way.

Coursera Update: Striking a Balance with Start Dates and Deadlines | Coursera Blog. (2013). Descargado 2018-06-13, de `https://blog.coursera.org/coursera-update-striking-a-balance-with-start/`

Daradoumis, T., Bassi, R., Xhafa, F., y Caballé, S. (2013). A review on massive e-learning (mooc) design, delivery and assessment. En *P2p, parallel, grid, cloud and internet computing (3pgcic), 2013 eighth international conference on* (pp. 208–213).

de Medeiros, A. K. A., Weijters, A. J., y van der Aalst, W. M. (2007). Genetic process mining: an experimental evaluation. *Data Mining and Knowledge Discovery*, *14*(2), 245–304.

de Waard, I., Abajian, S., Gallagher, M. S., Hogue, R., Keskin, N., Koutropoulos, A., y Rodriguez, O. C. (2011). Using mlearning and moocs to understand chaos, emergence, and complexity in education. *The International Review of Research in Open and Distributed Learning*, *12*(7), 94–115.

de Waard, I., Keskin, N. O., y Koutropoulos, A. (2016). Exploring future seamless learning research strands for massive open online courses. En *Human-computer interaction: Concepts, methodologies, tools, and applications* (pp. 2126–2140). IGI Global.

Elias, T. (2011). Learning analytics. *Learning*.

Emond, B., y Buffett, S. (2015). Analyzing student inquiry data using process discovery and sequence classification. *International Educational Data Mining Society*.

Eynon, R. (2013). *The rise of big data: what does it mean for education, technology, and media research?* Taylor & Francis.

Fritz, J. (2011). Classroom walls that talk: Using online course activity data of successful students to raise self-awareness of underperforming peers. *The Internet and Higher Education*, *14*(2), 89–97.

Günther, C. W., y Rozinat, A. (2012). Disco: Discover your processes. *BPM (Demos)*, *940*, 40–44.

Günther, C. W., y Van Der Aalst, W. M. (2007). Fuzzy mining–adaptive process simplification based on multi-perspective metrics. En *International conference on business process management* (pp. 328–343).

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., y Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, *2*(2-3), 107–124.

Ho, A. D., Chuang, I., Reich, J., Coleman, C. A., Whitehill, J., Northcutt, C. G., ... Petersen, R. (2015). Harvardx and mitx: Two years of open online courses fall 2012-summer 2014.

Jivet, I. (2016). *The learning tracker: a learner dashboard that encourages self-regulation in mooc learners* (Tesis de magíster). the Netherlands.

Jovanović, J., Gašević, D., Dawson, S., Pardo, A., Mirriahi, N., y cols. (2017). Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education*, *33*, 74–85.

Khalil, M., y Ebner, M. (2016). De-identification in learning analytics. *Journal of Learning Analytics*, *3*(1), 129–138.

Khodabandelou, G., Hug, C., Deneckere, R., y Salinesi, C. (2013). Process mining versus intention mining. En *Enterprise, business-process and information systems modeling* (pp. 466–480). Springer.

Kizilcec, R. F., Pérez-Sanagustín, M., y Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Computers & Education*, *104*, 18–33.

Kizilcec, R. F., y Schneider, E. (2015). Motivation as a lens to understand online learners: Toward data-driven design with the olei scale. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *22*(2), 6.

Kloos, C. D., Muñoz-Merino, P. J., Alario-Hoyos, C., Ayres, I. E., y Fernández-Panadero, C. (2015). Mixing and blending mooc technologies with face-to-face pedagogies. En *Global engineering education conference (educon), 2015 ieee* (pp. 967–971).

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S., y Hatala, M. (2015). Penetrating the black box of time-on-task estimation. En *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 184–193).

Larusson, J. A., y White, B. (2014). Introduction. En J. A. Larusson y B. White (Eds.), *Learning analytics: From research to practice* (pp. 1–12). New York, NY: Springer New York. Descargado de `https://doi.org/10.1007/978-1-4614-3305-7_1`

Leitner, P., Broos, T., y Ebner, M. (2018). Lessons learned when transferring learning analytics interventions across institutions. En *Companion proceedings 8th international conference on learning analytcis & knowledge.*

Littlejohn, A., y Milligan, C. (2015). Designing moocs for professional learners: Tools and patterns to encourage self-regulated learning. *eLearning Papers*, *42*.

Liu, Z., He, J., Xue, Y., Huang, Z., Li, M., y Du, Z. (2015). Modeling the learning behaviors of massive open online courses. En *Big data (big data), 2015 ieee international conference on* (pp. 2883–2885).

Lodge, J., y Lewis, M. (2012). Pigeon pecks and mouse clicks: Putting the learning back into learning analytics. *Future challenges, sustainable futures. Proceedings ascilite Wellington*, 560–564.

Lodge, J. M., y Corrin, L. (2017). What data and analytics can and do say about effective learning. *npj Science of Learning*, *2*(1), 5.

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., y Munoz-Gama, J. (2018). Mining theory-based patterns from big data: Identifying self-regulated learning strategies in massive open online courses. *Computers in Human Behavior*, *80*, 179–196.

Milrad, M., Wong, L.-H., Sharples, M., Hwang, G.-J., Looi, C.-K., y Ogata, H. (2013). Seamless learning: An international perspective on next-generation technology-enhanced learning.

Mukala, P., Buijs, J., y Van Der Aalst, W. (2015). *Uncovering learning patterns in a mooc through conformance alignments* (Inf. Téc.). Tech. rep., Eindhoven University of Technology, BPM Center Report BPM-15-09, BPMcenter. org.

Mukala, P., Buijs, J. C., Leemans, M., y van der Aalst, W. M. (2015). Learning analytics on coursera event data: A process mining approach. En *Simpda* (pp. 18–32).

Nesbit, J. C., Zhou, M., Xu, Y., y Winne, P. (2007). Advancing log analysis of student interactions with cognitive tools. En *12th biennial conference of the european association for research on learning and insruction (earli)* (pp. 2–20).

Pérez-Sanagustín, M., Hilliger, I., Alario-Hoyos, C., Kloos, C. D., y Rayyan, S. (2017). H-mooc framework: reusing moocs for hybrid education. *Journal of Computing in Higher Education*, *29*(1), 47–64.

Peterson, R. A. (1994). A meta-analysis of cronbach's coefficient alpha. *Journal of consumer research*, *21*(2), 381–391.

Pintrich, P. R., y cols. (1991). A manual for the use of the motivated strategies for learning questionnaire (mslq).

Rebuge, Á., y Ferreira, D. R. (2012). Business process analysis in healthcare environments: A methodology based on process mining. *Information systems*, *37*(2), 99–116.

Reimann, P., Frerejean, J., y Thompson, K. (2009). Using process mining to identify models of group decision making in chat data. En *Proceedings of the 9th international conference on computer supported collaborative learning-volume 1* (pp. 98–107).

Reimann, P., Markauskaite, L., y Bannert, M. (2014). e-research and learning theory: What do sequence and process mining methods contribute? *British Journal of Educational Technology*, *45*(3), 528–540.

Rigotti, T., Schyns, B., y Mohr, G. (2008). A short version of the occupational self-efficacy scale: Structural and construct validity across five countries. *Journal of Career Assessment*, *16*(2), 238–255.

Rojas, E., Munoz-Gama, J., Sepúlveda, M., y Capurro, D. (2016). Process mining in healthcare: A literature review. *Journal of biomedical informatics*, *61*, 224–236.

Romero, C., Cerezo, R., Bogarín, A., y Sánchez-Santillán, M. (2016). Educational process mining: A tutorial and case study using moodle data sets. *Data Mining and Learning Analytics: Applications in Educational Research*, *1*.

Romero, C., y Ventura, S. (2017). Educational data science in massive open online courses. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(1).

Rozinat, A., y Van der Aalst, W. M. (2008). Conformance checking of processes based on monitoring real behavior. *Information Systems*, *33*(1), 64–95.

Schoonenboom, J. (2007). *Trails in education: Technologies that support navigational learning*. Sense Publ.

Sharples, M., Kloos, C. D., Dimitriadis, Y., Garlatti, S., y Specht, M. (2015). Mobile and accessible learning for moocs. *Journal of interactive media in education*, *2015*(1).

Soffer, T., y Cohen, A. (2015). Implementation of tel aviv university moocs in academic curriculum: A pilot study. *The International Review of Research in Open and Distributed Learning*, *16*(1).

Song, M., y van der Aalst, W. M. (2007). Supporting process mining by showing events at a glance. En *Proceedings of the 17th annual workshop on information technologies and systems (wits)* (pp. 139–145).

Sonnenberg, C., y Bannert, M. (2015). Discovering the effects of metacognitive prompts on the sequential structure of srl-processes using process mining techniques. *Journal of Learning Analytics*, *2*(1), 72–100.

Stanford online coursework to be available on new open-source platform. (2013, junio). Descargado 2018-06-13, de `http://news.stanford.edu/news/2013/june/open-source-platform-061113.html`

Suriadi, S., Wynn, M. T., Ouyang, C., ter Hofstede, A. H., y van Dijk, N. J. (2013). Understanding process behaviours in a large insurance company in australia: A case study. En *International conference on advanced information systems engineering* (pp. 449–464).

Trcka, N., y Pechenizkiy, M. (2009). From local patterns to global models: Towards domain driven educational process mining. En *Intelligent systems design and applications, 2009. isda'09. ninth international conference on* (pp. 1114–1119).

Trevors, G., Feyzi-Behnagh, R., Azevedo, R., y Bouchet, F. (2016). Self-regulated learning processes vary as a function of epistemic beliefs and contexts: mixed method evidence from eye tracking and concurrent and retrospective reports. *Learning and Instruction*, *42*, 31–46.

Van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., ... others (2011). Process mining manifesto. En *International conference on business process management* (pp. 169–194).

Van der Aalst, W., Weijters, T., y Maruster, L. (2004a). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1128–1142.

Van der Aalst, W., Weijters, T., y Maruster, L. (2004b). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1128–1142.

Van der Aalst, W. M. (2011). *Process mining: Discovery, conformance and enhancement of business processes*. Springer.

Van der Aalst, W. M. (2016). *Process mining: data science in action*. Springer.

van der Aalst, W. M., De Beer, H., y van Dongen, B. F. (2005). Process mining and verification of properties: An approach based on temporal logic. En *Otm confederated international conferences."n the move to meaningful internet systems"* (pp. 130–147).

Van Dongen, B. F., de Medeiros, A. K. A., Verbeek, H., Weijters, A., y Van Der Aalst, W. M. (2005). The prom framework: A new era in process mining tool support. En *International conference on application and theory of petri nets* (pp. 444–454).

van Eck, M. L., Lu, X., Leemans, S. J., y van der Aalst, W. M. (2015). Pm ˆ 2: A process mining project methodology. En *International conference on advanced information systems engineering* (pp. 297–313).

Vidal, J. C., Vázquez-Barreiros, B., Lama, M., y Mucientes, M. (2016). Recompiling learning processes from event logs. *Knowledge-Based Systems*, *100*, 160–174.

Warr, P., y Downing, J. (2000). Learning strategies, learning anxiety and knowledge acquisition. *British journal of Psychology*, *91*(3), 311–333.

Washio, T., y Motoda, H. (2003). State of the art of graph-based data mining. *Acm Sigkdd Explorations Newsletter*, *5*(1), 59–68.

Weijters, A., y Ribeiro, J. (2011). Flexible heuristics miner (fhm). En *Computational intelligence and data mining (cidm), 2011 ieee symposium on* (pp. 310–317).

Weijters, A. J., y Van der Aalst, W. M. (2003). Rediscovering workflow models from event-based data using little thumb. *Integrated Computer-Aided Engineering*, *10*(2), 151–162.

Wong, L.-H. (2013). Analysis of students' after-school mobile-assisted artifact creation processes in a seamless language learning environment. *Journal of Educational Technology & Society*, *16*(2).

Wong, L.-H., y Looi, C.-K. (2011). What seams do we remove in mobile-assisted seamless learning? a critical review of the literature. *Computers & Education*, *57*(4), 2364–2381.

Zhang, J., Skryabin, M., y Song, X. (2016). Understanding the dynamics of mooc discussion forums with simulation investigation for empirical network analysis (siena). *Distance Education*, *37*(3), 270–286.

Zhang, Y. (2013). Benefiting from mooc. En *Edmedia: World conference on educational media and technology* (pp. 1372–1377).

ANEXOS

A. ANEXO A: CÓDIGO PARA VALIDACIÓN DE ALPHA DE CRONBACH

Este código corresponde a un *Jupyter Notebook* donde recalculamos los alpha de Cronbach para las distintas variables, considerando solo el conjunto de usuarios que pudieron ser asociados con su ID y finalmente fueron utilizados en el estudio.

```
import pandas as pd
import numpy as np
from scipy import stats

users_elec = pd.read_excel('reg_elec.xlsx')
users_orga = pd.read_excel('reg_orga.xlsx')
users_aula = pd.read_excel('reg_aula.xlsx')

users = pd.concat([users_elec, users_orga, users_aula],
    ignore_index=True)

data_encuesta = pd.read_excel('dat.xlsx')

user_list = users['userId'].tolist()
data_filtrada = data_encuesta[data_encuesta['
    ucchile_user_id'].isin(user_list)]

# Calculo de Alpha de Cronbach
#Obtenido de https://github.com/statsmodels/statsmodels/
    issues/1272

def CronbachAlpha(itemscores):
    itemscores = np.asarray(itemscores)
```

```
    itemvars = itemscores.var(axis=0, ddof=1)
    tscores = itemscores.sum(axis=1)
    nitems = itemscores.shape[1]
    calpha = nitems / float(nitems-1) * (1 - itemvars.sum()
        / float(tscores.var(ddof=1)))

    return calpha


vars_srl = ['Goal.Setting',
            'Strategic.Planning',
            'SelfEvaluation',
            'Task.Strategies',
            'Elaboration',
            'Help.Seeking']

nums = [4,4,3,6,3,4]


for i in range(0,6):
    var = vars_srl[i]
    list_var = []
    for j in range(1,nums[i]+1):
        list_var.append(var+'.'+str(j))
    itemscores = data_filtrada[list_var]
    c_alpha = CronbachAlpha(itemscores)
    print('C_Alpha_'+var+':_\n'+str(c_alpha))
```

```
# Datos referencia otra aplicacion
# Goal Setting: alpha = 0.86, M = 3.0, SD = 0.76
# Strategic Planning: alpha = 0.75, M = 3.1, SD = 0.65
# Self Evaluation: alpha = 0.80, M = 3.3, SD = 0.66
# Task Strategies: alpha = 0.78, M = 3.1, SD = 0.62
# Elaboration: alpha = 0.77, M = 3.3, SD = 0.64
# Help Seeking: alpha = 0.77, M = 2.6, SD = 0.79
# alpha = 0.92, M = 3.0, SD = 0.52

# Calculo de matriz de correlaciones

def getVariable(data, nombre_var, n_pregs):
    list_var = []
    for j in range(1,n_pregs+1):
        list_var.append(nombre_var+'.'+str(j))
    print(list_var)
    data_pregs = data[list_var].values()
    return np.mean(data_pregs, axis=1)

datos_srl = []

for i in range(0,6):
    var = vars_srl[i]
    list_var = []
    for j in range(1,nums[i]+1):
```

```
            list_var.append(var+'.'+str(j))
        datos_srl.append(data_filtrada[list_var].mean(axis = 1)
            .tolist())

df_srl = pd.DataFrame(np.transpose(np.array(datos_srl)),
    index = range(len(datos_srl[0])),columns = vars_srl)
```

B. ANEXO B: EJEMPLOS DE CÓDIGO PARA PROCESAMIENTO DE REGISTROS DE EVENTOS

B.1. Script para creación de un registro de eventos con datos de Coursera

```
#load user list and fine grained activities

import pandas as pd

course_progress = pd.read_csv('course_progress.csv', names
    = ['course_id', 'course_item_id','ucchile_user_id','
    course_progress_state_type_id','course_progress_ts'])
users = pd.read_csv('users.csv', usecols = ['
    ucchile_user_id','user_join_ts','country_cd'])
course_grades = pd.read_csv('course_grades.csv')

users_grades = pd.merge(users,course_grades[['
    ucchile_user_id','course_passing_state_id']], on = '
    ucchile_user_id', how = 'inner')

#%%Merge tables for create event log

log_fg = course_progress.merge(users_grades[['
    ucchile_user_id','country_cd','course_passing_state_id'
    ]], on = 'ucchile_user_id', how = 'inner')

#%%convert dates to seconds since epoch

from datetime import datetime
```

```
ts_list = log_fg['course_progress_ts'].values.tolist()
ts_list_datetime = []

for i in range(len(ts_list)):
    try:
        ts_aux = datetime.strptime(ts_list[i], '%Y-%m-%d %H
            :%M:%S.%f')
    except:
        ts_aux = datetime.strptime(ts_list[i], '%Y-%m-%d %H
            :%M:%S')
    ts_list_datetime.append(int(ts_aux.timestamp()))

join_ts_list = users['user_join_ts'].values.tolist()
join_ts_list_datetime = []

for i in range(len(join_ts_list)):
    try:
        join_ts_aux = datetime.strptime(join_ts_list[i], '%
            Y-%m-%d %H:%M:%S.%f')
    except:
        join_ts_aux = datetime.strptime(join_ts_list[i], '%
            Y-%m-%d %H:%M:%S')
    join_ts_list_datetime.append(int(join_ts_aux.timestamp
        ()))
```

```
dict_join_ts = dict(zip(users['ucchile_user_id'],
    join_ts_list_datetime))

#%%%sort the log per user and timestamp

log_fg['course_progress_ts'] = ts_list_datetime

log_fg = log_fg.sort(['ucchile_user_id','course_progress_ts
    '])

ts_list_datetime = log_fg['course_progress_ts'].tolist()

#%%%calculate session number
import numpy as np

nSession = 1
tThreshold = 60*45

session = np.zeros(len(ts_list_datetime))
session[0] = 1

userId = log_fg['ucchile_user_id'].values.tolist()

for i in range(1,len(log_fg)):

    if(userId[i] == userId[i-1]):
```

```
            if(ts_list_datetime[i] - ts_list_datetime[i-1] <
               tThreshold):
                session[i] = nSession
            else:
                nSession += 1
                session[i] = nSession
        else:
            nSession = 1
            session[i] = nSession

log_fg['session'] = session

#%%

course_items = pd.read_csv('course_items.csv')
course_item_types = pd.read_csv('course_item_types.csv')
course_progress_state_types = pd.read_csv('
   course_progress_state_types.csv')

log_fg = log_fg.merge(course_items[['course_item_id','
   course_item_type_id']], on = 'course_item_id')

#%%Convert log to SRL transition format

userId = log_fg['ucchile_user_id'].values.tolist()
session = log_fg['session'].values.tolist()
timestamp = log_fg['course_progress_ts'].values.tolist()
```

```
activityType = log_fg['course_item_type_id'].values.tolist
    ()
activityStatus = log_fg['course_progress_state_type_id'].
    values.tolist()
itemId = log_fg['course_item_id'].values.tolist()
country = log_fg['country_cd'].values.tolist()
courseGrade = log_fg['course_passing_state_id'].values.
    tolist()

count = dict(zip(course_items['course_item_id'].values.
    tolist(),list(np.zeros(len(course_items)))))

userId_log = [userId[0]]
session_log = [session[0]]
activity_log = ['BEGIN_SESSION']
timestamp_log = [timestamp[0]-1]
country_log = [country[0]]
courseGrade_log = [courseGrade[0]]

count[itemId[0]] = 1

for i in range(1,len(log_fg)):
    mismoUser = userId[i-1] == userId[i]
    mismaSesion = session[i-1] == session[i]

    aux = count[itemId[i]]
    count[itemId[i]] = aux + 1
```

```
repeating = (count[itemId[i]] > 1) \& (activityType[i]
   != 6)
evaluating = (count[itemId[i]] > 1) \& (activityType[i]
    == 6)
studyingTactics = ((activityType[i-1] != 6) \& (
   activityType[i] == 6))|((activityType[i] != 6) \& (
   activityType[i-1] == 6))
elaboration = (activityType[i] == 6) \& (activityStatus
   [i] == 2)
reading = (activityType[i] != 6) \& (activityStatus[i]
   == 2)


if(mismoUser \& mismaSesion):
    if(repeating):
        userId_log.append(userId[i])
        session_log.append(session[i])
        activity_log.append('REPEAT_L')
        timestamp_log.append(timestamp[i]+1)
        country_log.append(country[i])
        courseGrade_log.append(courseGrade[i])
    if(evaluating):
        userId_log.append(userId[i])
        session_log.append(session[i])
        activity_log.append('REPEAT_A')
        timestamp_log.append(timestamp[i]+1)
```

```
            country_log.append(country[i])
            courseGrade_log.append(courseGrade[i])
        if(reading):
            userId_log.append(userId[i])
            session_log.append(session[i])
            activity_log.append('LECTURE')
            timestamp_log.append(timestamp[i])
            country_log.append(country[i])
            courseGrade_log.append(courseGrade[i])
        if(elaboration):
            userId_log.append(userId[i])
            session_log.append(session[i])
            activity_log.append('ASSESS')
            timestamp_log.append(timestamp[i])
            country_log.append(country[i])
            courseGrade_log.append(courseGrade[i])
        if(studyingTactics):
            userId_log.append(userId[i])
            session_log.append(session[i])
            activity_log.append('PERFORM')
            timestamp_log.append(timestamp[i]-1)
            country_log.append(country[i])
            courseGrade_log.append(courseGrade[i])
    else:
        userId_log.append(userId[i-1])
        session_log.append(session[i-1])
        timestamp_log.append(timestamp[i-1]+1)
```

```
            activity_log.append('END SESSION')
            country_log.append(country[i-1])
            courseGrade_log.append(courseGrade[i-1])

            userId_log.append(userId[i])
            session_log.append(session[i])
            timestamp_log.append(timestamp[i])
            activity_log.append('BEGIN SESSION')
            country_log.append(country[i])
            courseGrade_log.append(courseGrade[i])

        if(not mismoUser):
            contador_elec = dict(zip(course_items['
                course_item_id'].values.tolist(),list(np.zeros(
                len(course_items)))))

    userId_log.append(userId[i])
    session_log.append(session[i])
    timestamp_log.append(timestamp[i]+1)
    activity_log.append('END SESSION')
    country_log.append(country[i])
    courseGrade_log.append(courseGrade[i])


    #%%
    dictLogSRL = {}
    dictLogSRL['userId'] = userId_log
```

```
dictLogSRL['session'] = session_log
dictLogSRL['timestamp'] = timestamp_log
dictLogSRL['activity'] = activity_log
dictLogSRL['country'] = country_log
dictLogSRL['courseGrade'] = courseGrade_log

logSRL = pd.DataFrame(dictLogSRL)

logSRL['courseId'] = course_progress['course_id'][0]

#%%
logSRL.to_csv('logSRL.csv', index = False)
```

B.2. Script para integrar variantes del proceso clasificadas con Disco

```
import pandas as pd

log_variantes = pd.read_csv('log_disco.csv', sep = ';')

#%%Contamos cuanta actividad corresponde a cada variante y
    cuantas veces se repite la variante

variantes = list(set(log_variantes['Variant'].tolist()))
count_variantes = log_variantes['Variant'].loc[
   log_variantes['Activity'] == 'Begin_Session'].
   value_counts()
```

```
actividades = ['Acomplete', 'Areview', 'Atry', 'Lbegin', '
    Lcomplete', 'Lreview']

act_variantes = pd.DataFrame(index = variantes, columns =
    actividades)
act_variantes['n'] = count_variantes
act_variantes.sort(['n'], ascending = False, inplace = True
    )
act_variantes['Case_ID'] = log_variantes['Case_ID'].tolist
    ()[0]

for variante in variantes:

    actividades_var = log_variantes['Activity'].loc[
        log_variantes['Variant'] == variante]

    for actividad in actividades:

        n_actividad = sum(actividades_var == actividad)/
            count_variantes[variante]
        act_variantes.ix[variante,actividad] = n_actividad

    case_id = log_variantes['Case_ID'].loc[log_variantes['
        Variant'] == variante].tolist()[0]
    act_variantes.ix[variante, 'Case_ID'] = case_id
```

```
#%%Importamos logs que tienen filtrados ciertos tipos de
    proceso para generar los motif

log_atry_lecture = pd.read_csv('log_disco_atry_lecture.csv'
    , sep = ';')
case_atry_lecture = set(log_atry_lecture['Case_ID'].tolist
    ())

log_lecture_acomplete = pd.read_csv('
    log_disco_lecture_acomplete.csv', sep = ';')
case_lecture_acomplete = set(log_lecture_acomplete['Case_ID
    '].tolist())

#%%Determinamos los motif de cada variante

act_variantes['motif'] = 'complex'

for variante in variantes:

    AC = act_variantes.ix[variante,'Acomplete'] != 0
    AR = act_variantes.ix[variante,'Areview'] != 0
    AT = act_variantes.ix[variante,'Atry'] != 0
    LB = act_variantes.ix[variante,'Lbegin'] != 0
    LC = act_variantes.ix[variante,'Lcomplete'] != 0
    LR = act_variantes.ix[variante,'Lreview'] != 0
```

```
    ATRY_L = act_variantes.ix[variante, 'Case_ID'] in
        case_atry_lecture
    L_ACOMPLETE = act_variantes.ix[variante, 'Case_ID'] in
        case_lecture_acomplete

# si no hay asessment
    if ~AC \& ~AR \& ~AT:
        act_variantes.ix[variante,'motif'] = 'only_lecture'
# si no hay lecture
    elif ~LB \& ~LC \& ~LR:
        act_variantes.ix[variante,'motif'] = 'only_
            assessment'
    elif ATRY_L \& ~L_ACOMPLETE:
        act_variantes.ix[variante,'motif'] = 'Atry_to_
            lecture'
    elif ~ATRY_L \& L_ACOMPLETE:
        act_variantes.ix[variante,'motif'] = 'lecture_to_
            Acomplete'

#%%incorporamos informacion de los motif al log de eventos

log_variantes['motif'] = 'complex'

for i in log_variantes.index.tolist():

    variante = log_variantes.ix[i, 'Variant']
```

```
    log_variantes.ix[i,'motif'] = act_variantes.ix[variante
        ,'motif']

#%%incorporamos informacion del usuario al log de eventos

log_variantes['User_ID'] = 'a'

for i in log_variantes.index.tolist():

    log_variantes.ix[i,'User_ID'] = log_variantes.ix[i,'
        Case_ID'].split('-')[0]

#%%exportamos el log con los motif

cols_exp = ['Case_ID', 'User_ID', 'Activity', 'Complete_
    Timestamp', 'course', 'completer','high_SRL', 'motif']

log_variantes.to_csv('log_motif.csv', columns = cols_exp,
    index = False, sep = ';')
```

C. ANEXO C: EJEMPLO DE CÓDIGO PARA CLUSTERIZACIÓN DE ESTUDIANTES Y ANÁLISIS

Este código corresponde a un *Jupyter Notebook* donde se clusterizan los estudiantes en base a su actividad en el curso, y también se aplican test estadísticos para comprobar diferencias entre los estudiantes de distintos clusters.

```
import pandas as pd
import numpy as np
from scipy import stats

sesiones = pd.read_csv('Log_definitivo_motifs.csv', sep = '
    ;')

users_elec = pd.read_excel('reg_elec.xlsx')
users_orga = pd.read_excel('reg_orga.xlsx')
users_aula = pd.read_excel('reg_aula.xlsx')

users = pd.concat([users_elec, users_orga, users_aula],
    ignore_index=True)

SRL_list = users['SRL'].tolist()
quartiles = np.percentile(SRL_list, np.arange(0, 100, 25))

SRL_quartile = []
for SRL_score in SRL_list:
    if SRL_score <= quartiles[1]:
        SRL_quartile.append(1)
    elif SRL_score <= quartiles[2]:
```

```
            SRL_quartile.append(2)
        elif SRL_score <= quartiles[3]:
            SRL_quartile.append(3)
        else:
            SRL_quartile.append(4)

users['SRL_q'] = SRL_quartile

sesiones = sesiones.merge(users[['userId', 'SRL_q']],
   left_on = 'User_ID', right_on = 'userId')

def z_test(x1,x2,n1,n2):
    p1 = x1/n1
    p2 = x2/n2
    p = (x1+x2)/(n1+n2)
    std = np.sqrt(p*(1-p)*(1/n1 + 1/n2))
    z = (p1-p2)/std
    p = 2*stats.norm.sf(np.abs(z))

    return np.round(z,3), p

def chisq_test(list_1,list_2):
    if len(list_1) == len(list_2):

        df = len(list_1)
        if np.sum(list_1) == np.sum(list_2):
            df -= 1
```

```
            k1 = np.sqrt(np.sum(list_2)/np.sum(list_1))
            k2 = np.sqrt(np.sum(list_2)/np.sum(list_1))
            chisq = np.sum((k1*list_1-k2*list_2)**2/(list_1+
                list_2))
            p = stats.chi2.sf(chisq,df)

            return np.round(chisq,3), p

        else:
            print('ERROR: No hay misma cantidad de categorias/
                bins')
            return 0, 0

motifs = ['only lecture',
          'Atry to lecture',
          'explore',
          'only asessment',
          'Lcomplete to Atry',
          'complex',
          'lecture to Acomplete']

## Clusters

#Crear tabla que cuenta cuantas sesiones de cada tipo hace
    cada usuario
```

```
user_list = users['userId'].tolist()
sesion_table = pd.DataFrame(np.zeros(shape=(len(user_list),
   len(motifs))), index = user_list, columns = motifs)

sesion_table['userId'] = user_list

for i in range(0,len(sesiones)):
    sesion_table.ix[sesiones.ix[i,'User ID'],sesiones.ix[i,
       'motif']] += 1

cluster_table = sesion_table.merge(users[['userId','SRL']],
   on = 'userId')

cluster_data = cluster_table[['only lecture',
          'Atry to lecture',
          'explore',
          'only asessment',
          'Lcomplete to Atry',
          'lecture to Acomplete','SRL']].values

get_ipython().magic('matplotlib inline')

import matplotlib
from sklearn import preprocessing as prep
from sklearn import decomposition as decomp
import matplotlib.pyplot as plt
```

```
scaled_data = prep.scale(cluster_data)

pca = decomp.PCA(n_components=2)
data_pca = pca.fit_transform(scaled_data)

from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_samples,
    silhouette_score

kmeans_dataset_2 = KMeans(n_clusters=2, random_state=0).fit
    (scaled_data)
c_dataset_2 = kmeans_dataset_2.labels_
plt.scatter(data_pca[:,0], data_pca[:,1], c = c_dataset_2)

silhouette_score(scaled_data, c_dataset_2)

kmeans_dataset_3 = KMeans(n_clusters=3, random_state=0).fit
    (scaled_data)
c_dataset_3 = kmeans_dataset_3.labels_

plt.scatter(data_pca[:,0], data_pca[:,1], c = c_dataset_3);

silhouette_score(scaled_data, c_dataset_3)
```

```
# Dada la metrica nos quedamos con 3 cluster, ya que tiene
    puntaje similar a 2 cluster, pero al agregar uno
    adicional cae abruptamente, lo que significa que mas
    clusters empeoran la clasificacion.

kmeans_dataset_4 = KMeans(n_clusters=4, random_state=0).fit
    (scaled_data)
c_dataset_4 = kmeans_dataset_4.labels_
plt.scatter(data_pca[:,0], data_pca[:,1], c = c_dataset_4)

silhouette_score(scaled_data, c_dataset_4)

kmeans_dataset_5 = KMeans(n_clusters=5, random_state=0).fit
    (scaled_data)
c_dataset_5 = kmeans_dataset_5.labels_
plt.scatter(data_pca[:,0], data_pca[:,1], c = c_dataset_5)

silhouette_score(scaled_data, c_dataset_5)

cluster_table['category'] = c_dataset_3
cluster_table.drop('userId',1, inplace = True)

labels = ['only_lecture',
          'Atry_to_lecture',
          'explore',
          'only_asessment',
          'Lcomplete_to_Atry',
```

```
            'lecture to Acomplete']

### Cluster 1 - K-Means - poca actividad
#
# Este cluster refleja comportamiento de estudiantes con
   poca actividad en el curso. Contiene 2821 estudiantes.

cluster_table[labels+['SRL']].loc[cluster_table['category'
   ]==0].mean()

cluster_table[labels+['SRL']].loc[cluster_table['category'
   ]==0].std()

plt.boxplot(cluster_table[labels].loc[cluster_table['
   category']==0].values);
plt.xticks([1,2,3,4,5,6],labels, rotation = 'vertical');
plt.ylim([0,80]);

plt.boxplot(cluster_table['SRL'].loc[cluster_table['
   category']==0].values);
plt.xticks([1],['SRL']);
plt.ylim([1,5]);


### Cluster 2 - K-means - orientados a assessments
#
```

```
# Este cluster refleja estudiantes con alta actividad,
    donde predomina actividad del tipo Atry to lecture, lo
    que evidencia que estos estudiantes muchas veces parten
    trabajando assessments para luego buscar en las lecturas
    , tienen casi nula actividad de assessment completado
    desde una lectura previa. Contiene 554 estudiantes.

cluster_table[labels+['SRL']].loc[cluster_table['category'
    ]==1].mean()

cluster_table[labels+['SRL']].loc[cluster_table['category'
    ]==1].std()

plt.boxplot(cluster_table[labels].loc[cluster_table['
    category']==1].values);
plt.xticks([1,2,3,4,5,6],labels, rotation = 'vertical');
plt.ylim([0,80]);

plt.boxplot(cluster_table['SRL'].loc[cluster_table['
    category']==1].values);
plt.xticks([1],['SRL']);
plt.ylim([1,5]);

### Cluster 3 - K-Means - orientados a lecturas
#
```

```
# En este cluster se detecta la actividad como esta
    intencionada en el curso, es decir multiples lecturas
    consecutivas, o pasar de una lectura a completar
    assessent, contiene solo 83 estudiantes.

cluster_table[labels+['SRL']].loc[cluster_table['category'
    ]==2].mean()

cluster_table[labels+['SRL']].loc[cluster_table['category'
    ]==2].std()

plt.boxplot(cluster_table[labels].loc[cluster_table['
    category']==2].values);
plt.xticks([1,2,3,4,5,6],labels, rotation = 'vertical');
plt.ylim([0,80]);

plt.boxplot(cluster_table['SRL'].loc[cluster_table['
    category']==2].values);
plt.xticks([1],['SRL']);
plt.ylim([1,5]);

# ### Info adicional
#
# Confirmamos tamanos de cada cluster

len(cluster_table.loc[cluster_table['category']==0])
len(cluster_table.loc[cluster_table['category']==1])
```

```
len(cluster_table.loc[cluster_table['category']==2])


# Ahora analizamos similitud de SRL entre cada cluster, se
    observa que niveles de SRL entre cluster 2 y 3 tienen
    media muy similar.

SRL_cluster1 = cluster_table['SRL'].loc[cluster_table['
    category']==0].values
SRL_cluster2 = cluster_table['SRL'].loc[cluster_table['
    category']==1].values
SRL_cluster3 = cluster_table['SRL'].loc[cluster_table['
    category']==2].values

stats.ttest_ind(SRL_cluster2,SRL_cluster3)
stats.ttest_ind(SRL_cluster1,SRL_cluster2)
stats.ttest_ind(SRL_cluster1,SRL_cluster3)

# Ahora usamos test KS para verificar si distribucion de
    SRL es distinta entre los pares de clusters, pero nos
    entrega como resultado que no podemos rechazar hipotesis
     nula de que son iguales las distribuciones de SRL.

stats.ks_2samp(SRL_cluster2,SRL_cluster3)
stats.ks_2samp(SRL_cluster1,SRL_cluster2)
stats.ks_2samp(SRL_cluster1,SRL_cluster3)
```

```
# ## Otros Clustering

# ### Gaussian Mixture
#
# Para mejorar el resultado obtenido con KMeans, usaremos
   GMM con 3 clusters.

from sklearn.mixture import GaussianMixture

gmm_fit_3 = GaussianMixture(n_components=3, random_state =
   20).fit(scaled_data)
c_dataset_3 = gmm_fit_3.predict(scaled_data)

plt.scatter(data_pca[:,0], data_pca[:,1], c = c_dataset_3);
silhouette_score(scaled_data, c_dataset_3)

# ### Aglomerativo

from sklearn.cluster import AgglomerativeClustering

c_dataset_3 = AgglomerativeClustering(n_clusters=3).fit(
   scaled_data).labels_

plt.scatter(data_pca[:,0], data_pca[:,1], c = c_dataset_3);

silhouette_score(scaled_data, c_dataset_3)
```

```
# ## Jerarquico

from scipy.cluster.hierarchy import dendrogram, linkage

Z = linkage(scaled_data, 'ward')

plt.figure(figsize=(15, 8))
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('sample index')
plt.ylabel('distance')
dendrogram(
    Z,
    leaf_rotation=90.,  # rotates the x axis labels
    leaf_font_size=12.,  # font size for the x axis labels
    truncate_mode='lastp',  # show only the last p merged
        clusters
    p=20,  # show only the last p merged clusters
)
plt.show()


# Se confirman k=3, como nro de clusters propuesto.

from scipy.cluster.hierarchy import fcluster

k = 3
```

```
c_dataset_3 = fcluster(Z, k, criterion='maxclust')

plt.scatter(data_pca[:,0], data_pca[:,1], c = c_dataset_3);

silhouette_score(scaled_data , c_dataset_3)

cluster_table['category'] = c_dataset_3

sum(c_dataset_3 == 2)


# ### Cluster 1 - aglomerativo - poca actividad
#
# Este cluster refleja comportamiento de estudiantes con
   poca actividad en el curso. Contiene 2619 estudiantes.

cluster_table[labels+['SRL']].loc[cluster_table['category'
   ]==1].mean()

cluster_table[labels+['SRL']].loc[cluster_table['category'
   ]==1].std()

plt.boxplot(cluster_table[labels].loc[cluster_table['
   category']==1].values);
plt.xticks([1,2,3,4,5,6],labels, rotation = 'vertical');
plt.ylim([0,80]);
```

```
plt.boxplot(cluster_table['SRL'].loc[cluster_table['
    category']==1].values);
plt.xticks([1],['SRL']);
plt.ylim([1,5]);

### Cluster 2 - aglomerativo - orientados a lecturas
#
# En este cluster se detecta la actividad como esta
    intencionada en el curso, es decir multiples lecturas
    consecutivas, o pasar de una lectura a completar
    assessent, contiene solo 127 estudiantes.

cluster_table[labels+['SRL']].loc[cluster_table['category'
    ]==2].mean()
cluster_table[labels+['SRL']].loc[cluster_table['category'
    ]==2].std()

plt.boxplot(cluster_table[labels].loc[cluster_table['
    category']==2].values);
plt.xticks([1,2,3,4,5,6],labels, rotation = 'vertical');
plt.ylim([0,80]);

plt.boxplot(cluster_table['SRL'].loc[cluster_table['
    category']==2].values);
plt.xticks([1],['SRL']);
plt.ylim([1,5]);
```

```
# ### Cluster 3 - aglomerativo - orientados a assessments
#
# Este cluster refleja estudiantes con alta actividad,
    donde predomina actividad del tipo Atry to lecture, lo
    que evidencia que estos estudiantes muchas veces parten
    trabajando assessments para luego buscar en las lecturas
    , tienen casi nula actividad de assessment completado
    desde una lectura previa. Contiene 712 estudiantes.

cluster_table[labels+['SRL']].loc[cluster_table['category'
    ]==3].mean()

cluster_table[labels+['SRL']].loc[cluster_table['category'
    ]==3].std()

plt.boxplot(cluster_table[labels].loc[cluster_table['
    category']==3].values);
plt.xticks([1,2,3,4,5,6],labels, rotation = 'vertical');
plt.ylim([0,80]);

plt.boxplot(cluster_table['SRL'].loc[cluster_table['
    category']==3].values);
plt.xticks([1],['SRL']);
plt.ylim([1,5]);

# Info adicional
# Confirmamos tamanos de cada cluster
```

```
len(cluster_table.loc[cluster_table['category']==1])
len(cluster_table.loc[cluster_table['category']==2])
len(cluster_table.loc[cluster_table['category']==3])

# Ahora analizamos similitud de SRL entre cada cluster, se
   observa que niveles de SRL entre cluster 2 y 3 tienen
   media muy similar.

SRL_cluster1 = cluster_table['SRL'].loc[cluster_table['
   category']==1].values
SRL_cluster2 = cluster_table['SRL'].loc[cluster_table['
   category']==2].values
SRL_cluster3 = cluster_table['SRL'].loc[cluster_table['
   category']==3].values

stats.ttest_ind(SRL_cluster2,SRL_cluster3)

stats.ttest_ind(SRL_cluster1,SRL_cluster2)

stats.ttest_ind(SRL_cluster1,SRL_cluster3)
```

```
# Estos resultados nos indican que se rechaza que el
    cluster 1 sea similar en SRL media a los otros dos
    cluster, es decir aquellos estudiantes que pertenecen al
     cluster de poca actividad, tienen SRL relativamente
    menor, de todas formas entre los otros dos cluster no se
     puede garantizar diferencias.

# Ahora usamos test KS para verificar si distribucion de
    SRL es distinta entre los pares de clusters, pero nos
    entrega como resultado que no podemos rechazar hipotesis
     nula de que son iguales las distribuciones de SRL.

stats.ks_2samp(SRL_cluster2,SRL_cluster3)

stats.ks_2samp(SRL_cluster1,SRL_cluster2)

stats.ks_2samp(SRL_cluster1,SRL_cluster3)

# Hay evidencia de que SRL en cluster 1 (poca actividad)
    tiene distribucion distinta al cluster 3 (actividad
    orientada a resolver assessments), no bien asi entre los
     otros pares.

# Ajuste a proporciones

# Ahora consideramos los valores representativos de cada
    cluster, pero normalizados
```

```
prop_1 = cluster_table[labels+['SRL']].loc[cluster_table['
    category']==1].mean().tolist()[0:6]
prop_1 = np.array(prop_1)/sum(prop_1)

prop_2 = cluster_table[labels+['SRL']].loc[cluster_table['
    category']==2].mean().tolist()[0:6]
prop_2 = np.array(prop_2)/sum(prop_2)

prop_3 = cluster_table[labels+['SRL']].loc[cluster_table['
    category']==3].mean().tolist()[0:6]
prop_3 = np.array(prop_3)/sum(prop_3)

props_srl = pd.DataFrame(data = np.transpose(np.array([
    prop_1, prop_2, prop_3])),
                         columns = ['Low Activity',
                                    'Lecture Oriented',
                                    'Assess Oriented'],
                         index = ['only lecture',
                                  'Atry to lecture',
                                  'explore',
                                  'only asessment',
                                  'Lcomplete to Atry',
                                  'lecture to Acomplete'])

props_srl
```

```
# Ahora realizamos un test Chi Cuadrado para comparar los
    vectores de proporciones en cada cluster.

# Low Activity vs Lecture Oriented

chisq, p = chisq_test(prop_1, prop_2)
print('chi_sq = ' + str(chisq))
print('p = ' + str(p))

# Low Activity vs Assess Oriented

chisq, p = chisq_test(prop_1, prop_3)
print('chi_sq = ' + str(chisq))
print('p = ' + str(p))

# Lecture Oriented vs Assess Oriented

chisq, p = chisq_test(prop_2, prop_3)
print('chi_sq = ' + str(chisq))
print('p = ' + str(p))

# Lo hacemos con los valores completos del n sesiones medio
    en cada cluster

n_1 = np.array(cluster_table[labels+['SRL']].loc[
    cluster_table['category']==1].mean().tolist()[0:6])
```

```
n_2 = np.array(cluster_table[labels+['SRL']].loc[
    cluster_table['category']==2].mean().tolist()[0:6])
n_3 = np.array(cluster_table[labels+['SRL']].loc[
    cluster_table['category']==3].mean().tolist()[0:6])

# Low Activity vs Lecture Oriented

chisq, p = chisq_test(n_1, n_2)
print('chi_sq = ' + str(chisq))
print('p = ' + str(p))

# Low Activity vs Assess Oriented

chisq, p = chisq_test(n_1, n_3)
print('chi_sq = ' + str(chisq))
print('p = ' + str(p))

# Lecture Oriented vs Assess Oriented

chisq, p = chisq_test(n_2, n_3)
print('chi_sq = ' + str(chisq))
print('p = ' + str(p))

# Se observa diferencia evidente entre cluster 1 y los
    otros 2, pero entre cluster 2 y 3 la diferencia no es
    tanta como para rechazar que sean iguales las medias.
```