UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Análisis y descubrimiento de patrones de comportamiento en estudiantes de cursos online

TRABAJO FIN DE MÁSTER

Máster Universitario en Gestión de la Información

Autor: Jorge Alberto Vázquez Mendoza
Tutor: Cèsar Ferri Ramírez
Tutor: Carlos Monserrat Aranda

Curso 2020-2021

Resum

Les tecnologies educatives emergents a nivell mundial han tingut un considerable impacte en l'aprenentatge de la població, entre les més destacades tenim els cursos *MOOC, els quals, s'estima que per a l'any 2024 representaren aproximadament 17 mil milions de dòlars, amb un creixement anual del 40%. Aquests cursos representen una manera accessible d'adquirir coneixement, ja que solament requereixen d'accés a internet, però, els seus participants tendeixen a adquirir actituds contraproduents a l'aprenentatge al no estar exposats a la pressió del compromís en ser cursos asincrònics, la qual cosa resulta en l'abandó o fracàs en el curs. En aquest treball es proposa un estudi exploratori que analitza els registres de navegació d'un conjunt d'estudiants del curs "*Basic *Spanish 1: *Getting *Started" per a identificar les variacions en el comportament de seguiment del material, així com patrons d'actituds contraproduents, principalment la procrastinació, basant-se en una segmentació de la mostra de dades per diferents característiques dels alumnes.

Paraules clau: *MOOC, mineria de processos, patrons de comportament, procrastinació

Resumen

Las tecnologías educativas emergentes a nivel mundial han tenido un considerable impacto en el aprendizaje de la población, entre las más destacadas tenemos los cursos MOOC, los cuales, se estima que para el año 2024 representaran aproximadamente 17 mil millones de dólares, con un crecimiento anual del 40 %. Estos cursos representan una manera accesible de adquirir conocimiento, ya que solamente requieren de acceso a internet, pero, sus participantes tienden a adquirir actitudes contraproducentes al aprendizaje al no estar expuestos a la presión del compromiso al ser cursos asincrónicos, lo que resulta en el abandono o fracaso en el curso. En este trabajo se propone un estudio exploratorio que analiza los registros de navegación de un conjunto de estudiantes del curso "Basic Spanish 1: Getting Started" para identificar las variaciones en el comportamiento de seguimiento del material, así como patrones de actitudes contraproducentes, principalmente la procrastinación, basándose en una segmentación de la muestra de datos por distintas características de los alumnos.

Palabras clave: MOOC, minería de procesos, patrones de comportamiento, procrastinación

Abstract

Emerging educational technologies worldwide have had a considerable impact on the learning of the population, among the most prominent we have the MOOC courses, which, it is estimated that by 2024 they will represent approximately 17 billion dollars, with an annual growth 40%. These courses represent an accessible way of acquiring knowledge, since they only require internet access, but their participants tend to acquire counterproductive attitudes to learning by not being exposed to the pressure of commitment as they are asynchronous courses, which results in abandonment. or failure in the course. In this paper an exploratory study is proposed that analyzes the navigation records of a group of students of the course "Basic Spanish 1: Getting Started" to identify variations in the behavior of following the material, as well as patterns of counterproductive attitudes, mainly the procrastination, based on a segmentation of the data sample by different characteristics of the students.

Key words: MOOC, process mining, behavior patterns, procrastination

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

Agradezco a mis padres Jorge y Blanca, quienes han sido la luz en mi camino durante toda mi vida, gracias por guiarme por el camino de la bondad, el respeto y la tolerancia hacia mis semejantes. Querido padre, gracias por ser el espejo en el que siempre me quise reflejar, y aunque te adelantaste en el camino por el que todos tarde o temprano tenemos que pasar, nunca me dejaste solo, siempre te sentí a mi lado, en los pequeños momentos en los que no podía continuar. Querida madre, gracias por tu valor al afrontar la vida por mi hermana y por mí. Gracias a los dos por demostrarme que el amor eterno sí existe.

Agradezco también a mi hermana Valeria, mi eterna compañera de vida, mi mejor amiga, gracias por todos esos momentos que pasamos creciendo juntos. Y gracias por los dos más grandes regalos que jamás pude recibir, mis sobrinos Jaime y Dulce María.

Quisiera extender un enorme agradecimiento a mis tutores Cèsar y Carlos, gracias por el apoyo brindado durante el desarrollo de este trabajo de fin de máster. Gracias por vuestro tiempo, dedicación y paciencia desde el inicio de este estudio. El cual es un escalón más en mi formación académica y profesional.

Agradezco también al área de sistemas de la información y las comunicaciones (ASIC), con especial mención a Ignacio Despujol, quien me ha brindado su ayuda durante el desarrollo de este trabajo de fin de máster, gracias por el apoyo y confianza, los cuales fueron clave para el éxito del estudio realizado.

Y como no, agradecer a mis amigos, quienes siempre han estado pendientes de mí, quienes siempre han sido ese hombro en el cual me puedo apoyar cuando el camino se torna difícil.

Jorge

CAPÍTULO 1
Introducción

El mundo cambia constantemente, y con él también lo hace nuestra sociedad, donde, la tecnología cada día se encuentra más involucrada en nuestras vidas, siendo así a tal punto, que muchas actividades diarias dependen en gran parte de dispositivos o herramientas de software para poder ser completadas. Esto, sumado a la situación sanitaria actual, nos ha obligado como sociedad a adaptarse a nuevas formas de convivir. Actividades tan básicas como hacer compras o participar en reuniones sociales dependen de medios tecnológicos para su acometido.

La educación con el pasar de los años, se ha visto impulsada por el uso de herramientas tecnológicas. De este modo, el uso de plataformas virtuales se ha convertido en una alternativa fiable para el aprendizaje y, con esto, su impacto en la sociedad y en la economía es cada vez más notorio.

Entre las diferentes tecnologías que han emergido en el ámbito educativo tenemos los llamados "Cursos Masivos Abierto en Línea" (MOOC por sus siglas en inglés Massive Open Online Courses), los cuales son material educativo diseñado para albergar a una gran cantidad de usuarios en cualquier parte del mundo teniendo, como requisito único tener acceso a internet. Son cursos asíncronos gratuitos que no tienen restricciones de acceso [1], pero existe una cantidad considerable de formas en que las instituciones y organizaciones los han monetizado [2].

El impacto de los MOOC en la economía según Technavio [3] ha llegado a un punto en el que se estima que entre los años 2020-2024 este mercado crecerá en aproximadamente 17 mil millones de dólares. Se prevé un progreso a una tasa compuesta anual de más del 40 % en el periodo antes mencionado. Entre los principales competidores para este mercado están Coursera Inc, edX Inc, FutureLearn Ltd., entre otros.

Los cursos MOOC y las plataformas que los contienen generan una cantidad descomunal de datos, entre los que se encuentra información demográfica de sus participantes, registros de navegación por el material educativo y registros de evaluación referentes a la completitud de las competencias que deben de ser adquiridas para completarlos. Dichos datos contienen un valor agregado de gran utilidad para las instituciones u organizaciones a las que pertenecen. Mediante el uso de diversas ciencias de la computación, es posible que este valor agregado permita mejorar el proceso de aprendizaje de sus estudiantes al identificar problemas o desafíos que deben ser solventados.

Varias son las ciencias de la computación que al trabajar con grandes cantidades de datos permiten encontrar en ellos un valor agregado beneficioso para sus usuarios u organizaciones a los que estos pertenecen. Entre esas ciencias se encuentran la minería de datos y la minería de procesos.

La correcta aplicación de estas ciencias sobre los datos generados por cursos MOOC permiten encontrar y afrontar varios problemas o desafíos a los que diariamente se enfrentan los estudiantes y tutores de los cursos. Gran parte de estos estudios se han centrado en predecir y/o prevenir el abandono de los participantes de los cursos. Otros estudios se han centrado en analizar patrones de comportamiento basándose en diversas teorías como es la autorregulación del aprendizaje. Sin embargo, pocos estudios se han centrado en determinar si ciertos comportamientos no recomendados, como es la procrastinación, tienen impacto alguno en el éxito o el fracaso de los participantes del curso.

Para el desarrollo de este trabajo de fin de máster del máster universitario en gestión de la información (MUGI), se propone un estudio exploratorio mediante la aplicación de técnicas de minería de procesos sobre los registros de eventos de un curso MOOC ofertado por la Universidad Politécnica de Valencia en busca de patrones de comportamiento contraproducentes y más específicamente identificar la procrastinación. También se analizará cómo estos patrones de comportamiento afectan al resultado del estudiante en el curso. A diferencia de otros trabajos en donde se aplica ciencias de la computación con el fin de determinar el posible resultado del estudiante en el curso, sea este positivo o negativo (finalización o abandono), en este trabajo la propuesta no se enfoca en el resultado del curso, sino en el proceso que lleva a ese resultado. Este estudio se realiza sobre el curso MOOC ofrecido por la Universidad Politécnica de Valencia titulado "Basic Spanish 1" en su primera edición del año 2020.

1.1 Motivación

Los motivos de este trabajo se pueden resumir en los siguientes puntos:

1. Aportar a la comunidad académica que fomenta el uso de plataformas virtuales como una nueva metodología de aprendizaje.

2. Aplicar los conocimientos adquiridos al haber cursado el Máster Universitario en Gestión de la Información.

3. Vincular los conocimientos adquiridos con mis estudios de grado en Ingeniería en Sistemas en la Universidad de Cuenca (Ecuador).

En la asignatura de explotación de datos masivos (EDM) se expone el valor agregado que pueden tener los datos generados por las diferentes plataformas utilizadas por las organizaciones e instituciones educativas. Este valor no es visible a primera vista, pero, al ser identificado, permite mejorar el funcionamiento de la organización. Relacionando esto a la necesidad de optimizar la funcionalidad de las plataformas educativas, se propone un análisis de los datos que se generan masivamente en las plataformas con el fin de afrontar uno de los desafíos del aprendizaje en línea menos estudiado: la procrastinación.

1.2 Objetivos

En este apartado, se exponen los objetivos, tanto el general como los específicos que fueron considerados primordiales para el proceso de investigación, estructuración y ejecución de este trabajo de fin de máster.

1.2.1. Objetivo general

Explorar patrones de procrastinación en el comportamiento de los participantes de un curso MOOC y determinar la relación que existe entre este comportamiento contraproducente y el éxito o fracaso de las personas inscritas en el curso.

1.2.2. Objetivos específicos

1. Conocer el trasfondo teórico de los conceptos y tecnologías a utilizar en el desarrollo de este trabajo.
2. Realizar un análisis bibliográfico mediante una revisión sistemática para identificar el estado del arte actual de la aplicación de minería de datos y minería de procesos en estudios enfocados a los desafíos del aprendizaje en línea.
3. Determinar la brecha de conocimiento existente en la aplicación de técnicas de minería de datos y minería de procesos para analizar los desafíos del aprendizaje en línea.
4. Proponer e implementar una solución efectiva en un caso de estudio mediante la ejecución de varios experimentos sobre los datos disponibles.
5. Basándose en los resultados obtenidos identificar las respuestas al objetivo general planteado.

1.3 Metodología

Para cumplir el objetivo principal de este documento es necesario plantear una metodología para la estructuración de las fases que se llevaron a cabo en la investigación de este trabajo de fin de máster. Para ello, los pasos a seguir son los siguientes:

- **Revisión sistemática de bibliografía.** Se procederá con la ejecución de una revisión sistemática de bibliografía para el apartado del estado del arte. De este modo, mediante la aplicación de una versión adaptada de la metodología de Barbara Kitchenham para revisiones sistemáticas de ingeniería de software [4] se identifican los vacíos en la investigación sobre los desafíos del aprendizaje autodirigido en cursos MOOC.
- **Análisis del problema.** Definición formal de la problemática a solventar con este trabajo de fin de máster, así como la identificación de las posibles soluciones tecnológicas que permitan solucionarla.
- **Propuesta de una solución efectiva al problema.** Se propondrá un conjunto de pasos a seguir para solucionar y/o solventar la problemática expuesta en el análisis del problema.
- **Extracción y limpieza de datos.** Se procederá con la extracción y limpieza de los registros de navegación de la plataforma MOOC, con el fin de generar un registro de eventos limpio sobre el cual implementar las tecnológicas seleccionadas para resolver el problema planteado.
- **Análisis de los modelos identificados y redacción de resultados.** Para finalizar, se procederá con el análisis e interpretación de los modelos de procesos generados por los registros de eventos analizados con el propósito de identificar los patrones

de comportamiento requeridos y las variaciones de estos basándose en los registros clasificados.

1.3.1. Estructura de la investigación

En esta sección se describen brevemente los apartados desarrollados para el presente documento.

1. **Capítulo 2. Marco teórico – Aprendizaje automático:** Se introducen conceptos básicos sobre las ciencias de la computación utilizadas durante el desarrollo de este estudio.
2. **Capítulo 3. Marco teórico:** Se presentan conceptos básicos sobre qué es un MOOC, así como el aprendizaje autodirigido en entornos digitales.
3. **Capítulo 4. Análisis sistemático de bibliografía:** Se realiza una revisión sistemática de literatura en donde se analiza la existencia de estudios previos que manejan los desafíos del aprendizaje en línea en un contexto de cursos MOOC.
4. **Capítulo 5. Solución propuesta:** Se desarrolla y describe detalladamente una solución en base a los resultados obtenidos en el capítulo 4. En esta solución se especifican los pasos y requerimientos necesarios para su adecuada aplicación.
5. **Capítulo 6. Caso de estudio:** Se expone el desarrollo de un caso de estudio donde se analiza y aplica a detalle la solución propuesta en el capítulo 5. Para ello, se utiliza el curso MOOC ofertado por la Universidad Politécnica de Valencia "Basic Spanish 1" en su edición 2020.
6. **Capítulo 7. Resultados:** Se presentan los resultados obtenidos a partir del caso de estudio presentado en el capítulo 6.
7. **Capítulo 8. Conclusiones:** Se detallan los resultados y conclusiones obtenidos durante la ejecución del caso de estudio, así como las posibles líneas de trabajos futuros que podrían surgir a partir de este trabajo y los trabajos derivados.

CAPÍTULO 2
Marco teórico – Aprendizaje automático

En este capítulo se introduce el trasfondo teórico relacionado a las ciencias de la computación consideradas para dar solución al problema propuesto en este estudio. En donde, se aborda los principales aspectos y conceptos teóricos que se requiere introducir, siendo estos:

1. **Minería de datos.** La sección 2.1 define y detalla la minería de datos con sus respectivas técnicas y características de cada una de ellas.

2. **Minería de procesos.** La sección 2.2 aborda los conceptos relacionados con la minería de procesos, así como sus respectivas técnicas.

2.1 Minería de datos

La minería de datos (DM por sus siglas en inglés Data Mining) según Rizvi [5] es la extracción y descubrimiento de conocimiento a partir de un gran conjunto de datos. Por otra parte, Gorunescu [6] define la minería de datos como la exploración y análisis automático o semiautomático de un gran conjunto de datos con el fin de descubrir patrones significativos, mientras que, Kantardzic [7] define a la minería de datos como un proceso que se encarga de la búsqueda de información nueva, valiosa y no trivial de un gran conjunto de datos.

Con base en lo anteriormente expuesto, podemos definir a la minería de datos como un proceso exploratorio que a partir de un gran conjunto de datos puede extraer información, conocimiento o patrones significativos que sean de utilidad para el usuario o entorno al que pertenecen.

Rizvi [5] señala que este proceso consta de seis pasos (Figura 2.1), siendo éstos: limpieza, extracción, selección, transformación, minería de datos propiamente dicha y evaluación de patrones.

Figura 2.1: Pasos de la minería de datos

Fuente: Elaboración propia

2.1.1. Técnicas de minería de datos

Existen diversas clasificaciones de las técnicas de DM basándose en el punto de vista de los autores. Rizvi [5] identifica a las técnicas de DM como clasificación, clusterización, regresión, reglas de asociación y redes neuronales.

Por otra parte, Pujari [8] en su libro "Data mining techniques", destaca que las técnicas de DM pueden definirse en grupos de diferentes maneras, pero que la mayoría de las técnicas forman parte de más de un grupo. Por ejemplo, basándose en el objetivo de DM, sus técnicas pueden ser predictivas o descriptivas. Por otra parte, basándose en la intervención del usuario, las técnicas pueden ser modelos de verificación (guiadas por el usuario) o modelos de descubrimiento (no guiadas por el usuario o automáticas).

Las técnicas que Pujari [8] describe son las técnicas de asociación, clasificación, patrones secuenciales, reglas de asociación, clusterización, redes neuronales, algoritmos genéticos y máquinas de vectores de soporte (SVM). También indica la existencia de minerías derivadas de la MD como es el Web Mining y el Text Mining, pero estas no serán consideradas en esta revisión, puesto que no forman parte del enfoque del estudio.

Basándose en lo dicho por los autores citados anteriormente, en este trabajo se consideran las siguientes técnicas de minería de datos: técnicas de clasificación, clusterización, regresión, reglas de asociación, redes neuronales, algoritmos genéticos y SVM, las cuales serán descritas a detalle a continuación.

Clasificación

En minería de datos, para Kesavaraj [9], la clasificación es asignar elementos de una colección a categorías o clases destino. Su objetivo es predecir con precisión la clase destino a la cual pertenece cada dato. En esta definición se identifican varios modelos de clasificación, entre los que tenemos: Clasificación binaria, Árboles de decisión, Métodos basados en reglas, Redes neuronales, Redes bayesianas y Máquinas de vectores de soporte (SVM).

Por otra parte, para Rizvi [5], la clasificación es una técnica de minería de datos que se caracteriza por arrancar con un conjunto de información clasificada previamente, la cual denominaremos datos de entrenamiento. Estos datos de entrenamiento son usados por un algoritmo de clasificación para el primer paso denominado "entrenamiento", el cual, obtiene como resultado un modelo denominado "clasificador". Si este es lo suficientemente preciso será capaz de clasificar nuevos registros de datos de la misma categoría de los datos de entrenamiento, pero no vistos durante la fase de entrenamiento. Para el

autor existen cinco modelos de clasificación: árboles de decisión, redes bayesianas, redes neuronales y máquinas de vectores de soporte.

Con base en las definiciones de clasificación anteriormente expuestas, la categorización de sus modelos que será utilizada en este trabajo será la siguiente:

1. Clasificación binaria
2. Árboles de decisión
3. Redes bayesianas
4. Máquinas de vectores de soporte

No se considera a las redes neuronales, puesto que, por su complejidad, serán tratadas como una categoría independiente más adelante.

Clusterización

La clusterización, para Gorunescu [6], es una técnica de minería de datos en la que se parte de un conjunto, el cual se divide en varios subconjuntos basándose en una o varias similitudes predeterminadas. De este modo, podría ser considerado como una especie de clasificación, pero con la diferencia de que no se busca determinar una característica desconocida para asignar un elemento a un conjunto, sino que la pertenencia del elemento a éste viene dado por sus características ya conocidas. El éxito del proceso de clusterización es determinado si la similitud entre elementos similares y diferencia con elementos pertenecientes a otros conjuntos se maximiza. Para ello se utilizan medidas de similitud entre las que tenemos la distancia euclidiana, la medida de Tanimoto, el coeficiente de Pearson, etc.

Para Kantardzic [7], clusterización son un conjunto de metodologías para clasificación automática en donde, de un conjunto de datos, se obtiene una serie de subconjuntos o conglomerados en donde sus elementos comparten características similares y difieren de otros subconjuntos por las mismas características. El análisis de estos subconjuntos permite obtener una descripción generalizada de cada uno de ellos, la cual es fundamental para un estudio a profundidad de los conglomerados.

Por otra parte, para Rizvi [5], la clusterización es una técnica de minería de datos que clasifica objetos en clases similares mediante técnicas de agrupación. De este modo se pueden descubrir patrones de distribución de estos por sus características. Para Rizvi los principales métodos de clusterización son:

1. Métodos de partición
2. Métodos jerárquicos de aglomeración
3. Métodos basados en densidad
4. Métodos basados en cuadrículas
5. Métodos basados en modelos

Los otros autores no especifican un conjunto específico de métodos de clusterización en sus definiciones.

Regresión

La regresión, también llamada correlación, para Gorunescu [6] es una técnica de minería de datos que, con base en un modelo matemático, mediante una ecuación de regresión establece la conexión entre los valores de una variable de tipo resultado (mejor conocida como variable dependiente) y un conjunto de variables predictoras (también llamadas variables independientes).

La definición establecida por otros autores es similar a la del anteriormente citado. Por ejemplo, Rizvi [5] define a la regresión como una técnica que se utiliza para modelar la relación que existe entre un conjunto de variables por determinar o dependientes con un conjunto de variables conocidas o independientes.

Las diferentes categorías de técnicas de regresión varían según el autor. Por ejemplo, para Rizvi [5] las técnicas de regresión pueden ser lineales o no lineales a la vez que pueden ser de una sola variable o de múltiples variables.

Por otra parte, para Olson [10], la regresión puede ser regresión normal o logística dependiendo si se pretende con ella predecir valores o clasificar valores respectivamente. Para autores como Chapple [11], la regresión se clasifica en dos grupos: La regresión lineal y la regresión múltiple. Esta última se subdivide en cuatro clases: regresión múltiple estándar, regresión múltiple escalonada, regresión jerárquica y regresión de Setwise.

Reglas de asociación

Para Gorunescu [6], las reglas de asociación son técnicas de minería de datos no supervisadas que se manejan como implicaciones de la forma $X \rightarrow Y$, en donde tanto X como Y son elementos o conjuntos de elementos distintos. X o los elementos de X son conocidos como el antecedente de la regla mientras que, Y o los elementos de Y son conocidos como el consecuente de la regla. Dicho de otra forma, la parte antecedente de la regla es aquella que debe de satisfacerse con el fin de que la parte consecuente sea verdadera.

Gorunescu [6] señala que los algoritmos más comunes para la resolución de reglas de asociación son el algoritmo a priori, el algoritmo de patrón de crecimiento frecuente (FP-growth) y el algoritmo ECLAT (Equivalence Class Clustering and Bottom-up Lattice Traversal).

A su vez, Kantardzic [7] define a las reglas de asociación como una de las principales técnicas de minería de datos y la forma más común de descubrimiento de patrones locales en sistemas no supervisados. Esta metodología se caracteriza porque su proceso de extracción de datos tiende a hacer un análisis exhaustivo de toda la base de datos con el fin de encontrar patrones que no eran evidentes a simple vista o que no eran de conocimiento para el usuario. Aunque este descubrimiento puede resultar contraproducente, puesto que el usuario puede llegar a abrumarse ante la gran cantidad de información nueva. El análisis de usabilidad resulta difícil y requiere de demasiado tiempo.

Kantardzic [7] también especifica un conjunto de algoritmos que considera más comunes, siendo estos: el algoritmo a priori, el método de crecimiento frecuente (FP GROWTH METHOD) y el método de clasificación asociativa.

Finalmente, Rizvi [5] enfatiza en que las reglas de asociación se utilizan para encontrar patrones en grandes conjuntos de datos con el fin de ser de ayuda en la toma de decisiones. Dichos patrones son llamados reglas cuyo valor de confianza debe de ser menor a uno para que sea útil. Ahora bien, la cantidad de reglas de asociación que se generan a partir de un gran conjunto de datos suele ser muy grande y un alto porcentaje de ellas tiene poco o ningún valor.

Aunque Rizvi [5] no especifica qué algoritmos de reglas de asociación considera más comunes, sí hace énfasis en una clasificación de los tipos de reglas de asociación, los cuales son reglas de asociación multi nivel, multi dimensionales y cuantitativas.

Redes neuronales artificiales

Las redes neuronales artificiales son sistemas de procesamiento distribuido paralelo masivo compuestos por unidades de procesamiento simples. Tienen la capacidad de aprender conocimiento experimental expresado a través de las conexiones entre unidades de procesamiento y ponerlo a disponibilidad para su uso [7]. Estas requieren de una cantidad masiva de datos para su entrenamiento dependiendo de la cantidad de unidades de procesamiento que disponga y, consecuentemente, de la cantidad de parámetros a aprender. Las principales características de las redes neuronales artificiales son:

1. **No linealidad.** Todas sus unidades de procesamiento interactúan en la red en paralelo, aunque internamente pueden generarse grupos de unidades que funcionan linealmente.
2. **Aprendizaje por el ejemplo.** Una red neuronal artificial modifica la interacción entre sus unidades basándose en grandes conjuntos de datos de entrenamiento.
3. **Adaptabilidad.** Una red neuronal artificial tiene la capacidad de adaptar los pesos de sus interconexiones a los cambios del entorno.
4. **Respuesta probatoria.** Una red neuronal artificial puede proporcionar a la par con la información de una clase o decisión específica la confianza de ésta, la cual puede ser utilizada para rechazar datos ambiguos con el fin de mejorar el rendimiento de la red.
5. **Tolerancia a fallos.** Una red neuronal artificial puede ser inherentemente tolerante a fallos o ser capaz de realizar cálculos computacionales robustos sin que su rendimiento se degrade. Sin embargo, las redes neuronales son vulnerables ante los ataques adversariales, los cuales consisten en introducir en la información de entrada de una red neuronal datos incorrectos con el objetivo de producir resultados erróneos en la red. [12]

Las unidades de procesamiento más simples son conocidas como neuronas artificiales, las cuales son la unidad fundamental de procesamiento en cualquier red neuronal artificial. Éstas están conformadas por tres partes: enlaces de conexión, sumador de señales y la función de activación.

Para Kantardzic [7] existen dos tipos principales de redes neuronales, las de propagación y las recurrentes. Las primeras se propagan en una sola dirección desde las entradas a las salidas sin retroalimentación mientras que, las segundas, se caracterizan por tener un comportamiento en bucle en el que interviene un componente de sincronización.

Gorunescu [6], define a las redes neuronales artificiales como sistemas de procesamiento de información adaptativos no programados, los cuales dependen de ejemplos y se comportan como cajas negras. Es decir, la forma en que procesan la información no es explícita y en este caso a la unidad de procesamiento se la denomina perceptrón.

Para Gorunescu [6] existen tres formas de clasificar las redes neuronales artificiales: por su arquitectura, por su funcionamiento y por su forma de aprendizaje. Por su arquitectura las redes neuronales artificiales se clasifican en tres categorías: redes de alimentación directa en una sola capa, redes de alimentación directa multicapa y las redes

recurrentes. Por su funcionamiento se clasifican en redes de propagación hacia adelante y propagación hacia atrás (llamadas anteriormente como propagación y recurrentes). Finalmente, por su forma de aprendizaje tenemos las redes de aprendizaje supervisado y aprendizaje no supervisado.

2.2 Minería de procesos

La minería de procesos para Dos Santos [13] es una disciplina enfocada en comprender los procesos en tiempo real basándose en la información recopilada por los diferentes sistemas de información. Con esta es posible conocer, optimizar y mejorar los diferentes procesos cubiertos por los sistemas.

A su vez, Bogarin [14] describe a la minería de procesos como una metodología relativamente nueva que surge con el objetivo de desarrollar técnicas destinadas a la extracción del conocimiento relacionado con los procesos presentes en registros de eventos con el fin de descubrir, monitorear y mejorar los procesos en diferentes dominios.

Por otra parte, para Van der Aalst [15] la minería de procesos es una disciplina que surge ante el crecimiento exponencial de datos de eventos generados por los sistemas de información. Esto abre la puerta a nuevas oportunidades de análisis de los procesos de tal manera que permita a los usuarios analizar eventos reales extrayendo el conocimiento de los registros disponibles en los sistemas de información actuales.

A base de los tres puntos de vista anteriormente expuestos, podemos decir que la minería de procesos es una disciplina que, basándose en los registros de eventos generados por los sistemas de información, está en capacidad de extraer conocimiento de estos con el fin de descubrir, monitorear o mejorar los procesos cubiertos por el sistema y extraer conocimiento útil de estos.

Según dos Santos [13] existen 3 técnicas principales para la aplicación de minería de procesos, las cuales se detallan brevemente a continuación.

2.2.1. Descubrimiento de procesos

Las técnicas de descubrimiento de procesos construyen un modelo de comportamiento basándose en un registro de eventos, no parten de ningún modelo base y se utilizan principalmente en el control de flujo [14]. Entre los principales algoritmos se encuentran:

1. Alpha Algorithm
2. Heuristic Mining
3. Multiphase Mining
4. Fuzzy Mining
5. Genetic Mining
6. Region Miner
7. Integer Linear Programming Miner
8. Declarative Miner2

2.2.2. Conformidad de procesos

La conformidad de procesos o verificación de conformidad parte de un modelo base y modela el comportamiento real basándose en un registro de eventos. Sirve para encontrar concordancias o discrepancias entre lo esperado y lo observado [14]. Según Van der Aalst [15] existen cuatro dimensiones de calidad para el análisis del enfoque de conformidad, las cuales son:

1. **Idoneidad.** En donde el modelo de procesos es capaz de visualizar la mayor parte del comportamiento presente en el registro de eventos. La idoneidad del modelo de procesos es catalogada como perfecta cuando permite visualizar todas las variantes del proceso.
2. **Simplicidad.** En donde mientras más simple sea el modelo de procesos resultante mejor puede explicar el comportamiento del o de los procesos analizados.
3. **Precisión.** Un modelo de procesos es preciso si no expresa en sus resultados excesiva cantidad de variaciones. Si un modelo no presenta precisión se considera desajustado, que significa que el modelo de procesos generaliza excesivamente el comportamiento que se pretende analizar, lo que resulta en presentar comportamientos no relacionados con la realidad.
4. **Generalización.** Contrario a la dimensión anterior, el modelo no debe sobre ajustar sus resultados, es decir, no debe restringir la información presentada a unos pocos comportamientos específicos.

Por otra parte, el mismo autor [15] señala que la conformidad de procesos también se centra en dos enfoques de los modelos base, siendo estos los siguientes:

1. **Modelo descriptivo.** El cual simplemente evidencia el proceso ideal que debe de presentar el modelo de procesos resultante del análisis. En este escenario el resultado se valida por su alineación con el modelo base [16].
2. **Modelo normativo.** El cual describe el proceso estándar que debe de ser cubierto por el registro de eventos. La inconformidad del modelo resultante con respecto al modelo base puede evidenciar fraudes, ineficiencias y procedimientos desactualizados o mal diseñados [16].

2.2.3. Mejora de procesos

Las técnicas de mejora de procesos, al igual que la anterior, parten de un modelo inicial y, en relación con el modelo generado con los registros del comportamiento real, genera un nuevo modelo sobre el cual se puede identificar problemas en el flujo con el fin de optimizarlo [13]. El mismo autor hace referencia a los diferentes enfoques que puede tomar un estudio que aplique minería de procesos en el contexto de mejora de estos. Estos enfoques son los siguientes:

1. **Ampliación del modelo.** La posibilidad de enriquecer el modelo de procesos con otras perspectivas, correlacionado consigo mismo. De esta forma, la información adicional del registro de eventos brinda nuevos enfoques.
2. **Técnicas predictivas.** Combinación de enfoques predictivos con las técnicas de minería de procesos como los árboles de decisión, sistemas de recomendación, redes

neuronales, etc. Esto, con el fin de aportar a la información descubierta predicciones de la duración del evento, predicción del comportamiento basándose en información diferencial, etc.

3. **Perspectiva organizacional.** Combinando la mejora de procesos con el análisis de redes sociales se puede enriquecer el modelo organizacional, mejorarlo y aportar al mismo desde diferentes puntos de vista.

4. **Deriva o cambio en el proceso.** Identificar periodos de tiempo en los que se produce o se está produciendo un cambio en el proceso, localizar el punto del cambio y la razón de éste.

5. **Minería de decisiones.** Basándose en el modelo de proceso generado, es posible tomar decisiones para mejorar el modelo de negocios. Por ejemplo, sincronizar el proceso de venta con el proceso de fabricación para cubrir las existencias siempre sin la necesidad de sobrecargar al proceso de fabricación.

6. **Asignación de recursos.** A través de la mejora de procesos es posible identificar ausencia o sobrecargo de recursos tanto humanos como materiales o tecnológicos. La reasignación de estos recursos conforme se demuestre la necesidad del proceso optimiza el mismo.

CAPÍTULO 3
Marco teórico – MOOC y aprendizaje autodirigido

En este capítulo se exponen los conceptos relacionados con cursos MOOC y aprendizaje autodirigido. El objetivo es introducir al lector las bases teóricas necesarias para comprender los aspectos tratados posteriormente en este estudio. A grandes rasgos, los conceptos abordados a continuación son:

- **Cursos MOOC.** La sección 3.1 introduce los conceptos relacionados con los cursos masivos abiertos en línea (MOOC).
- **Aprendizaje autodirigido.** La sección 3.2 presenta los conceptos relacionados con el aprendizaje autodirigido en entornos en línea, y los desafíos que representa.

3.1 Cursos MOOC

Los cursos masivos, abiertos y en línea (MOOC por sus siglas en ingles Massive Open Online Courses) son, como su nombre indica, cursos que se encuentran disponibles en internet cuyo objetivo principal es permitir a usuarios de todo el mundo acceder a conocimiento que sea de su interés. Aunque existe una amplia variedad de definiciones sobre lo que es un curso MOOC, el análisis realizado por Hidalgo et al. [17] recopila varias de ellas. De estas se puede decir que los cursos MOOC son cursos diseñados para una gran cantidad de participantes, los cuales pueden acceder al mismo desde cualquier lugar en cualquier momento siempre que se disponga de una conexión a internet. Su característica más significativa es el libre acceso al contenido, a partir del cual el participante podrá decidir la naturaleza de uso de este material acorde a sus necesidades.

Aunque los cursos MOOC inicialmente son concebidos como cursos abiertos, las entidades que los ofertan han ideado diferentes formas de monetización de su contenido mediante la implementación de diferentes modelos de negocio [2]. Entre los más destacados se encuentran:

1. Modelo de monetización de certificados
2. Modelo de monetización de evaluaciones
3. Modelo de suscripciones

3.1.1. Características de los cursos MOOC

Las principales características de los cursos MOOC vienen dadas por sus iniciales, es decir, se trata de cursos masivos, abiertos y en línea. Ahora bien, en la literatura, varios autores definen un conjunto de características de las cuales son parte los MOOC. Kennedy [18] en la revisión sistemática realizada para el periodo 2009–2012 las define de la siguiente manera:

1. **Apertura.** La apertura del flujo de información es una característica vital para este tipo de sistemas complejos, los MOOC se caracterizan por usar software libre, porque su contenido sea libre y por brindar apertura al conocimiento para cualquier participante.

2. **Barreras de persistencia.** Los MOOC se caracterizan por tener una baja retención de estudiantes, en donde, el abandono se da principalmente en participantes cuyo idioma nativo no es el mismo en el cual es impartido el material educativo

3. **Modelos.** En rasgos generales, los cursos MOOC se identifican en dos grandes categorías, siendo estos los cMOOC los cuales se basan en un modelo conectivista (enfocados a estrategias de aprendizaje permanente) y los xMOOC que se basan en un modelo de conductismo cognitivo (enfocados en periodos de tiempo concretos y que su comportamiento se asemeja a una clase tradicional automatizada y distribuida masivamente)

3.1.2. Clasificación de los cursos MOOC

En la literatura se puede encontrar una gran variedad de clasificaciones para los cursos MOOC. Como se expuso anteriormente una clasificación general de los cursos MOOC puede ser los xMOOC y los cMOOC [18]. Por otra parte, según Atiaja et al. [19] los cursos MOOC se pueden clasificar basándose en su alcance y en su funcionalidad para el aprendizaje:

1. **Clasificación por alcance**

   a) **BMOOCs** (Large Scale Open Online Courses) Los cuales son cursos limitados a un número de participantes, usualmente se trata de cursos desarrollados para ser usados en una cátedra en particular.

   b) **DMOOCs** (Distributed Open Collaborative Courses) Se trata de cursos cuyo material es distribuido a un conjunto de participantes pertenecientes a diferentes instituciones educativas, pero quien administra dicho material pertenece solo a una de ellas.

   c) **MOORs** (Massive Open Online Research) El curso mezcla material educativo típico de un MOOC (videos, lecturas) con proyectos de investigación.

   d) **LOOC** (Little Open Online Course) Son cursos limitados a un pequeño conjunto de participantes que deben de pagar una tarifa de acceso.

   e) **SPOCs** (Small Online and Private Courses) Son cursos con un número limitado de participantes, comúnmente dirigidos a ser utilizados como complemento a una clase tradicional.

   f) **SMOCs** (Synchronous Massive Online Courses) Se trata de cursos síncronos en los que el material del curso es transmitido como si se tratará de una clase tradicional, por lo que se requiere acceder al mismo en un horario específico.

Aunque el concepto de los SMOCS se asemeja al de los xMOOC su principal diferencia es que los SMOCS no son "abiertos". Su acceso se limita a un grupo concreto que disponga del tiempo y los recursos requeridos por el curso [20].

2. **Clasificación por funcionalidad**

   a) **Transfer MOOC** (MOOC Transferred) Se trata de cursos no concebidos inicialmente como MOOC, pero adaptados a esta tecnología con el fin de ser distribuidos masivamente.

   b) **Made MOOC** (MOOC specifically created) Son cursos los cuales fueron concebidos desde un inicio como MOOCs, en los cuales se utilizó laboratorios especializados en el desarrollo de su contenido.

   c) **Synch MOOC** (Synchronized MOOC) Son cursos que cuentan con una fecha de inicio, fecha de fin y fechas límite para cumplir con tareas y evaluaciones.

   d) **Asynchronic MOOC** (MOOC asynchronous) Los cuales son cursos que no cuentan con fecha de inicio ni fin

   e) **Adaptative MOOC** (MOOC adapted) Son cursos que utilizan algoritmos para adaptar la experiencia de seguimiento del participante conforme este recorre el material

   f) **Group MOOC** (Group MOOC) Los cuales son cursos en los que los participantes trabajan en pequeños grupos que se mantendrán hasta finalizarlo.

   g) **Connectivism MOOC** (MOOC connection) Son cursos que siguen una filosofía de formación flexible basada en la interacción y el trabajo en grupo.

   h) **Mini MOOC** Son cursos cortos y con contenido específico que se utilizan como complemento en cátedras de diferentes universidades.

Pero, Atiaja et al. [19] recalca que muchas de estas variaciones de los cursos MOOC incumplen con sus principales características, aunque la disposición y estructura del mismo se asemeja en gran parte a lo que es un curso MOOC.

3.2 Aprendizaje autodirigido

El aprendizaje autodirigido es la capacidad de un alumno de guiar su propio proceso de aprendizaje [21]. Según lo dicho por Song [21], este tipo de aprendizaje se enfoca en dos perspectivas: la perspectiva del proceso y la perspectiva de los atributos personales.

3.2.1. Perspectiva del proceso

Song [21] afirma que el proceso de aprendizaje autodirigido se centra en tres grandes etapas primarias: la planificación, el monitoreo y la evaluación. Estas, brevemente, pueden ser descritas como:

1. **Planificación.** Es la capacidad que tiene el estudiante de marcar su propio ritmo de seguimiento del curso. En este caso, al no tener un tutor presencial quien coordine sus horarios y tiempos de aprendizaje como en una clase tradicional, es potestad del estudiante cómo se organiza para cumplir con los objetivos.

2. **Monitoreo.** Es la capacidad autocrítica del alumno de identificar si el conocimiento que se le está transmitiendo está siendo asimilado adecuadamente. Al no tener un tutor presencial que pueda determinar mediante gestos o actitud la comprensión del tema tratado, la responsabilidad de esta etapa recae sobre el estudiante.

3. **Evaluación.** El autor [21] indica que varios estudios han llegado a la conclusión de que la clase en línea requiere de mayor tiempo y esfuerzo que una clase tradicional para el tutor del tema tratado. Por ello se requiere que el estudiante pueda evaluarse a sí mismo y a sus compañeros.

3.2.2. Perspectiva de atributos personales

La perspectiva de atributos personales se centra en tres grupos, siendo estos el uso de recursos, el uso de estrategias y la motivación. A continuación se describen las perspectivas a la par con los problemas a los que se exponen:

1. **Recursos.** Son componentes del entorno de aprendizaje que pueden adoptar diferentes formas, como es el caso de los recursos humanos y los recursos de información. En un entorno de aprendizaje en línea el uso adecuado de recursos implica mayor participación en dinámicas de intercambio de opiniones, como es el caso de foros de discusión, y el uso adecuado del material disponible. El principal problema de la perspectiva de recursos es que al momento que el aprendizaje se torna asíncrono entre el estudiante y el tutor, si el primero requiere asesoría con respecto al material educativo, los posibles lapsos de tiempo entre pregunta y respuesta pueden generar incertidumbre a dicho estudiante.
2. **Estrategias.** El aprendizaje exitoso en todos los entornos de aprendizaje implica el uso de estrategias efectivas. Una buena estrategia implica disciplina por parte del estudiante y en el contexto de la educación en línea es primordial para los estudiantes comprender por su cuenta el material educativo. El no tener una estrategia efectiva para aprender en un entorno virtual es el principal problema al que se enfrenta esta perspectiva, puesto que puede afectar al interés por parte del estudiante. Además, hay que sumar a esto la poca interacción social resultante de no tener un entorno de cobertura de dudas inmediato que malas interpretaciones de la información recibida.
3. **Motivación.** La motivación a seguir un curso en línea está relacionada directamente con el interés personal que tiene el estudiante sobre el material a aprender. Mientras más alto sea el interés más atención prestará al contenido y su aprendizaje será más efectivo. Al no tener la motivación suficiente para prestar atención al material educativo impartido, es posible que no se adquiera el conocimiento requerido para culminar exitosamente el proceso de aprendizaje. Por otra parte, la facilidad de posponer el aprendizaje asíncrono se relaciona directamente con la procrastinación por parte de los estudiantes. Esto lo diferencia de una clase presencial, en donde resulta imposible para el estudiante posponer indefinidamente la continuidad del seguimiento del material educativo.

Al tener definidas las perspectivas del aprendizaje autodirigido, es necesario identificar los principales desafíos a los que se enfrentan tanto estudiantes como tutores en el contexto del aprendizaje en línea.

3.2.3. Desafíos del aprendizaje autodirigido

Basándose en las perspectivas expuestas anteriormente, tanto desde el punto de vista de proceso de aprendizaje como las perspectivas de atributos personales se han identificado cinco desafíos a los que se enfrentan tanto los estudiantes como tutores en los entornos de aprendizaje en línea. Enfocándonos en un contexto de MOOC, los desafíos son los siguientes:

1. **Abandono.** Relacionado directamente con el problema de las perspectivas de motivación y estrategia, en donde la desmotivación y la falta de estrategia de aprendizaje pueden desembocar en el abandono prematuro del curso. Esto es grave, ya que según Mrhar [22] la tasa de deserción en el contexto de los MOOC es aproximadamente del 90 % de los inscritos.

2. **Autorregulación del aprendizaje.** Desafío relacionado con el problema de la perspectiva de estrategia. Formalmente hablando, a la capacidad de aprender de forma autónoma de un estudiante se le conoce como capacidad de Autorregulación del aprendizaje [23]. Aquellos estudiantes que autorregulan mejor su aprendizaje obtienen mejores resultados en los entornos virtuales educativos [21], por lo que el principal desafío es cubrir las diferencias que implica esta capacidad en los diversos estudiantes.

3. **Interés o Atención.** Relacionado también con el problema de la perspectiva de motivación, para los tutores es un desafío mantener el interés de los estudiantes que no se sienten atraídos por el tema tratado. Por ello, es posible que el estudiante recorra el material educativo sin sumergirse en su contenido. Por ejemplo, que reproduzca el vídeo mientras navega en redes sociales [21].

4. **Procrastinación.** Relacionado con el problema de la perspectiva de motivación, al ser un entorno de aprendizaje asíncrono, un estudiante poco motivado tiende a postergar el seguimiento de las actividades el mayor tiempo posible [21].

5. **Participación.** Finalmente, el último desafío es la participación, que se encuentra vinculado a los problemas expuestos para los contextos de estrategia y motivación. Puesto que, un estudiante con una buena estrategia de aprendizaje autodirigido está consciente de la importancia de la participación en actividades de intercambio de ideas, como es el caso de los foros de discusión. De este modo aunque no tenga una estrategia eficiente, si tiene interés por el tema tratado participará. En caso contrario no lo hará [21].

CAPÍTULO 4
Análisis sistemático de bibliografía

En este capítulo se realiza un estudio secundario denominado revisión sistemática de bibliografía. El objetivo de éste es determinar el vacío en la investigación de los desafíos a los que se enfrentan los estudiantes y tutores en el aprendizaje autodirigido en el contexto de cursos MOOC mediante la aplicación de las ciencias de la computación denominadas minería de datos y minería de procesos.

4.1 Revisión sistemática de literatura

Una revisión sistemática de literatura es un método científico caracterizado por ser repetible y replicable y que se utiliza para recolectar información de un tema en específico. Para ello, se requiere seguir una metodología conformada por un conjunto de pasos con el fin de obtener una base sólida de información y, a la vez, determinar adecuadamente posibles vacíos en el campo analizado. Para el caso específico de este estudio se sigue una versión adaptada de la metodología propuesta por Barbara Kitchenham [4] la cual, en rasgos generales, cubre tres pasos:

1. Planificación de la revisión.
2. Ejecución de la revisión.
3. Informe de la revisión.

4.1.1. Metodología de investigación

Planificación de la revisión

La etapa de planificación está compuesta de tres pasos:

1. Preguntas de investigación.
2. Estrategia de búsqueda.
3. Criterios de selección de estudios primarios.

Preguntas de investigación. El objetivo de esta revisión sistemática es identificar y determinar qué desafíos del aprendizaje autodirigido han sido más comúnmente objeto de estudio mediante la aplicación de técnicas de minería de datos y minería de procesos. Para cumplir con este objetivo, se plantea la siguiente pregunta de investigación: ¿Qué

desafíos del aprendizaje autodirigido en cursos MOOC han sido objeto de estudio mediante la aplicación de técnicas de minería de datos y/o minería de procesos? A partir de esta pregunta se definen las siguientes sub-preguntas de investigación:

RQ1. ¿Qué técnicas de minería de datos o minería de procesos son utilizadas en los cursos MOOC?

RQ2. ¿En qué desafíos del aprendizaje autodirigido se han aplicado técnicas de minería de datos o minería de procesos?

RQ3. ¿Cómo se aborda la investigación de los desafíos del aprendizaje autodirigido en cursos MOOC mediante la aplicación de técnicas de minería de datos o minería de procesos?

Estrategias de búsqueda: La estrategia de búsqueda se realiza mediante la ejecución de búsquedas automáticas. Esta estrategia requiere una combinación de palabras clave basadas en las preguntas y sub-preguntas de investigación. Para esta revisión sistemática, se considerarán las cuatro principales bibliotecas digitales de ciencias de la computación: IEEE Explore, SCOPUS, ACM y Springer Link.

A la combinación de palabras clave se le conoce como cadena de búsqueda. La formación de esta cadena se encuentra detallada en la tabla 4.1

Concepto	Sub-cadena	Conector	Términos alternativos
Cursos masivos abiertos en línea	Massive Open Online Courses	AND	MOOC*
Minería de datos	Data Mining	OR	
Minería de procesos	Process Mining	OR	
Desafíos del aprendizaje autodirigido	Challenge*	AND	Dropout Self-Regulated Learning Distract* Procrastination Participation
Estudio	Study		Study

Tabla 4.1: Identificación de palabras clave para la cadena de búsqueda

Basándose en la tabla 4.1 se procede con la generación de la cadena de búsqueda, la cual se utiliza en las bibliotecas digitales para obtener los diferentes documentos académicos (artículos, conferencias, etc.) que coincidan con las palabras clave. La cadena de búsqueda utilizada es: *("Massive Open Online Courses" OR MOOC*) AND ((Data OR Process) AND Mining) AND (Challenge* OR dropout OR "Self-Regulated Learning" OR distract* OR Procrastination OR Participation) AND (Study).*

Criterios de selección de estudios primarios: Los criterios de extracción (EC por sus siglas en inglés Extraction Criteria) son usados para responder a las preguntas de investigación. La estrategia utilizada garantizará el cumplimiento de los criterios de extracción y ayudará a la clasificación de los estudios. En la Tabla 3.2 se muestran los criterios de extracción definidos. Los criterios que no tienen opciones son criterios abiertos, cuyos valores serán identificados durante el proceso de lectura a profundidad de los artículos analizados.

Criterio	Nombre	Opciones
RQ1.¿Qué técnicas de minería de datos o minería de procesos son utilizadas en los cursos MOOC?		
EC01	Ciencia de la computación	Minería de datos Minería de procesos
EC02	Técnica de minería de datos	Clasificación Clusterización Regresión Reglas de asociación Redes neuronales Support Vector Machine
EC03	Técnica de minería de procesos	Descubrimiento de procesos Conformidad de procesos Regresión Reglas de asociación Mejora de procesos
EC04	Algoritmo de minería de procesos	
EC05	Dimensiones de calidad del modelo de conformidad	Aptitud Simplicidad Precisión Generalización
EC06	Enfoque de conformidad	Modelo descriptivo Modelo normativo
EC07	Enfoque de mejora de procesos	Ampliación del modelo Técnicas predictivas Perspectiva organizacional Deriva o cambio en el proceso Minería de decisiones Asignación de recursos.
EC08	Herramientas adicionales	
RQ2. ¿En qué desafíos del aprendizaje autodirigido se han aplicado técnicas de minería de datos o minería de procesos?		
EC09	Desafíos del proceso de aprendizaje autodirigido en entornos online	Abandono Procrastinación Interés Autorregulación Participación
EC10	Tipo de datos analizados	Información demográfica Información de evaluación Registros de navegación
RQ3. ¿Cómo se aborda la investigación de los desafíos del aprendizaje autodirigido en cursos MOOC mediante la aplicación de técnicas de minería de datos o minería de procesos?		
EC11	Fase	Análisis Diseño Implementación Testeo

EC12	Tipo de validación	Prueba de conceptos Encuesta Experimento Prototipo Caso de estudio
EC13	Alcance del enfoque	Industria Academia
EC14	Metodología	Nueva Extensión
EC15	Año de publicación	
EC16	País	

Tabla 4.2: Criterios de extracción

Ejecución de la revisión

La etapa de ejecución está compuesta por tres pasos:

1. Selección de estudios primarios.
2. Evaluación de calidad.
3. Análisis y síntesis.

Selección de estudios primarios. Los artículos resultantes de las búsquedas en las bibliotecas digitales pasan por un proceso de selección. Este proceso de selección consta de tres etapas:

1. Evaluación de criterios de inclusión y exclusión.
2. Selección de artículos por título.
3. Selección de artículos por lectura a profundidad.

Se consideraron artículos a partir del año 2015, los criterios de inclusión y exclusión utilizados fueron:

- Criterios de inclusión
  - Estudios que utilizan técnicas de minería de datos en datos generados por plataformas MOOC.
  - Estudios que utilizan técnicas de minería de procesos en datos generados por plataformas MOOC.
  - Estudios que hayan sido realizados desde el año 2015 hasta hoy.
- Criterios de exclusión
  - Artículos introductorios de ediciones especiales: revistas, libros, conferencias
  - Artículos duplicados en diferentes fuentes

- Artículos cortos de menos de cinco páginas
- Artículos escritos en otras lenguas diferentes de inglés

Una vez finalizado el proceso de selección de estudios primarios se obtuvo los resultados expuestos en la tabla 4.3

Librería	Resultados de búsqueda	Seleccionados (Etapas 1 y 2)	Aceptados (Etapa 3)	Porcentaje del total aceptado
IEEE Explore	22	20	12	12,24 %
Scopus	71	44	14	14,29 %
ACM	524	217	32	32,65 %
Springer Link	818	312	40	40,82 %

Tabla 4.3: Resumen de los resultados de búsqueda

El total de artículos seleccionados en las etapas uno y dos es de 593, de los cuales, el total de artículos seleccionados es de 98. Estos representan el 6,83 % de la muestra inicial y el 16,53 % de los artículos seleccionados en las primeras dos etapas.

Evaluación de calidad: En este estudio, el aspecto considerado para evaluar la calidad de los estudios aceptados es el número de citas que el artículo ha tenido. Para realizar esta clasificación se determinó tres categorías: Alto, Medio y Bajo de acuerdo al contador de citaciones proporcionado por la plataforma Google Scholar. Esta decisión se tomó porque la relevancia de un artículo se refleja en la cantidad de veces que este ha sido referenciado en otros estudios. En la tabla 4.4 se encuentran expuestos los resultados de la evaluación de calidad.

Categoría/Puntuación	Criterio	# Artículos
Alto	Artículos con más de cinco citaciones	48
Medio	Artículos con una a cinco citaciones	30
Bajo	Artículos sin citaciones	20

Tabla 4.4: Evaluación de calidad

Análisis y síntesis. Para el análisis y síntesis de esta revisión sistemática se utilizó la herramienta Atlas.ti 9, la cual permite etiquetar el contenido de un artículo con el fin de determinar su enfoque basándose en los criterios de extracción abordados en la Tabla 3.2.

Figura 4.1: Logo de Atlas TI

Fuente: https://atlasti.com/

Basándose en el análisis ejecutado en Atlas.ti se obtuvo los resultados estadísticos individuales expuestos en la Tabla 4.5 para cada criterio de extracción en relación con los artículos que los abordan. Se omiten los criterios EC05, EC06 y EC07 porque no se encontraron artículos relacionados a estos criterios.

Criterio	Opciones	# Estudios	Porcentaje	Suma
EC01	Minería de datos	91	90,10 %	101
	Minería de procesos	10	9,90 %	
EC02	Clasificación	35	27,56 %	127
	Clusterización	22	17,32 %	
	Regresión	33	25,98 %	
	Reglas de asociación	6	4,72 %	
	Redes neuronales	19	14,96 %	
	Máquinas de vectores de soporte	12	9,45 %	
EC03	Descubrimiento de procesos	9	100,00 %	9
	Conformidad de procesos	0	0,00 %	
	Mejora de procesos	0	0,00 %	
EC04	Celonis Algorithm	1	7,14 %	14
	Disco Algorithm	2	14,29 %	
	Epistemic Network Analysis	1	7,14 %	
	EM algorithm	1	7,14 %	
	Fuzzy Miner Algorithm	3	21,43 %	
	Heuristic algorithm	1	7,14 %	
	Inductive Miner	2	14,29 %	
	PM2 method	1	7,14 %	
	SECPI	1	7,14 %	
	cSPADE	1	7,14 %	
EC08	Cuestionario	7	21,88 %	32
	Encuesta	2	6,25 %	
	Estadística	11	34,38 %	
	Inteligencia Artificial	1	3,13 %	
	Learning Analytics	4	12,50 %	
	Machine Learning	4	12,50 %	
	Sequence Mining	3	9,38 %	
EC09	Abandono	39	33,62 %	116
	Procrastinación	1	0,86 %	
	Interés	17	14,66 %	
	Autorregulación del aprendizaje	10	8,62 %	
	Participación	49	42,24 %	
EC10	Información demográfica	22	17,46 %	126
	Información de evaluación	14	11,11 %	
	Registros de navegación	90	71,43 %	
EC11	Análisis	91	88,35 %	103
	Diseño	3	2,91 %	
	Implementación	3	2,91 %	
	Pruebas	6	5,83 %	
EC12	Prueba de conceptos	0	0,00 %	104
	Encuesta	0	0,00 %	
	Experimento	55	52,88 %	
	Prototipo	2	1,92 %	
	Caso de estudio	47	45,19 %	
	Otros	0	0,00 %	
EC13	Industria	0	0,00 %	98
	Academia	98	100,00 %	
EC14	Nueva	98	100,00 %	98
	Extensión	0	0,00 %	

Tabla 4.5: Resultados obtenidos por criterio de extracción

Informe de la revisión

En esta sección los resultados obtenidos en la revisión sistemática son expuestos. Los resultados son divididos en tres categorías:

1. Síntesis y principales resultados.
2. Análisis de metadatos.
3. Análisis de contenido.

Síntesis y principales resultados: Los principales resultados obtenidos se presentan gráficamente en este apartado. Estos resultados son representados en un plano de coordenadas (X, Y) junto a una descripción detallada de lo observado.

En la Figura 4.2 el eje de las abscisas representa a la técnica de minería de datos mientras que el eje de las ordenadas representa el desafío del aprendizaje autodirigido abordado mediante dicha técnica. El tamaño de los círculos representa la cantidad de estudios que aplican la técnica de minería de datos con respecto al desafío de aprendizaje.

Se observa que los desafíos del aprendizaje más estudiados con técnicas de minería de datos son el abandono y la participación. En estos, el abandono es mayoritariamente estudiado utilizando técnicas de clasificación y redes neuronales, mientras que, la participación es, en su mayoría, estudiada aplicando técnicas de regresión y clasificación.

Por otra parte, los desafíos del aprendizaje menos estudiados son la procrastinación y la autorregulación, en donde destaca el hecho de que la procrastinación ha sido considerada sólo en un estudio de la muestra.

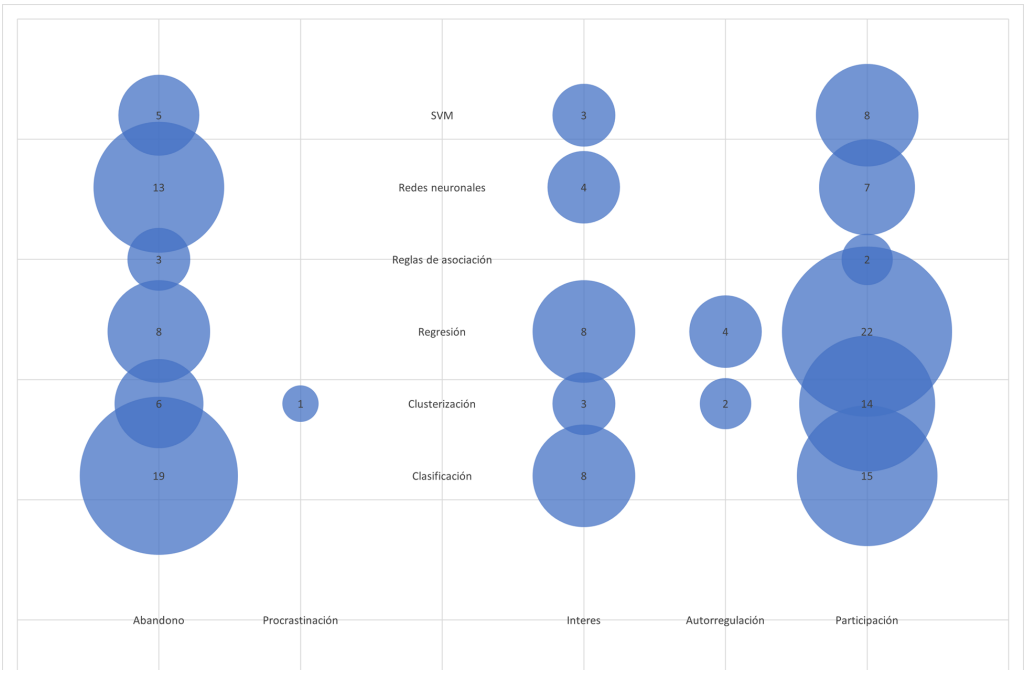

Figura 4.2: Relación establecida entre las técnicas de minería de datos y los desafíos del aprendizaje

En la Figura 4.3, en el eje de las abscisas se encuentra los desafíos del aprendizaje mientras que en el eje de las ordenadas tenemos los diferentes tipos de datos identificados, al igual que en la Figura 1 el tamaño de los círculos representa la cantidad de estudios que relacionan los valores de ambos ejes.

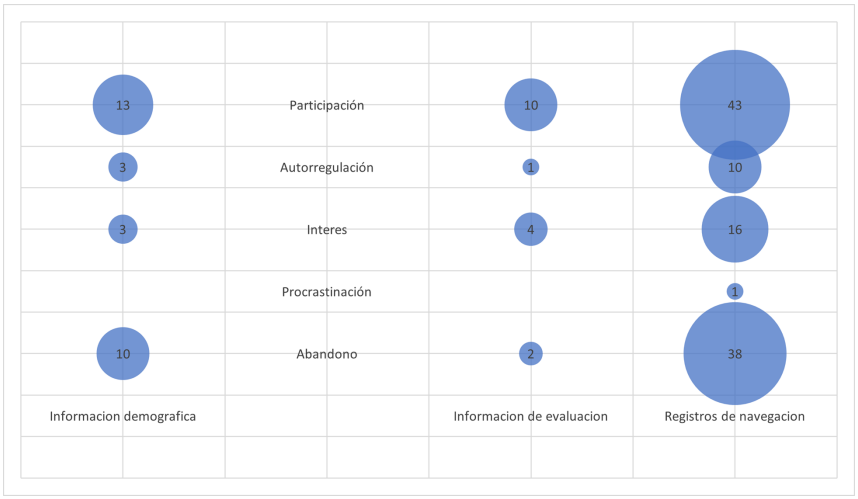

Figura 4.3: Relación establecida entre los desafíos del aprendizaje autodirigido con respecto al tipo de datos generados por los cursos MOOC.

En la Figura 4.3 se observa que los registros de navegación en la plataforma son los datos predilectos para analizar los desafíos del aprendizaje, seguidos por la información demográfica y, en menor medida, la información de evaluación. Los registros de navegación se utilizaron para todos los desafíos, siendo el desafío que más los utilizó la participación seguido del abandono. Por otra parte, la información demográfica se utiliza principalmente en estudios que analizan la participación y el abandono, finalizando con la información de evaluación, que fue utilizada en estudios de participación.

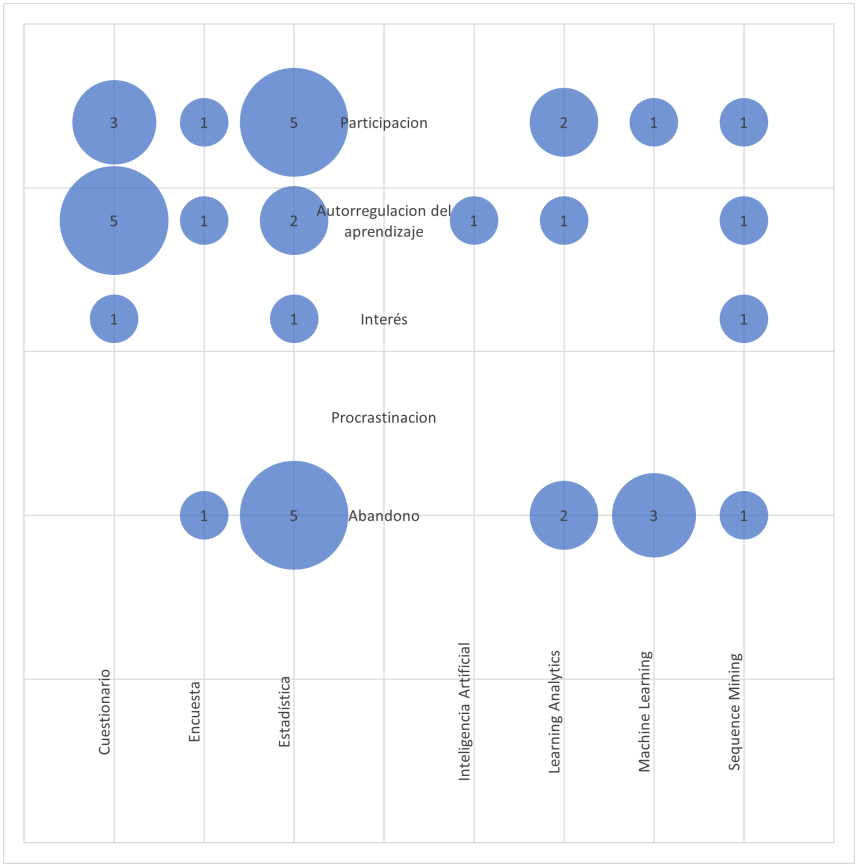

Figura 4.4: Relación establecida entre el desafío del aprendizaje y las herramientas adicionales identificadas

Finalmente, en la Figura 4.4 observamos la relación de los desafíos del aprendizaje autodirigido con respecto a las herramientas adicionales utilizadas en los estudios. Estas herramientas son recursos adicionales que utilizan los autores de los artículos aceptados para obtener mejores resultados en sus estudios.

Este estudio identificó siete herramientas adicionales utilizadas por los investigadores que aplican técnicas de minería de datos o minería de procesos durante la ejecución de sus estudios. De las siete técnicas identificadas, el uso de estadística antes y/o después de la aplicación de la minería de datos es común en estudios que se enfocan en el abandono o en la participación. Por otra parte, los cuestionarios al igual que la inteligencia artificial son más utilizados en estudios que se enfocan en la autorregulación del aprendizaje.

Análisis de metadatos: Para el análisis de metadatos se consideran los criterios de extracción EC15 y EC16 de la tabla 3.2, los cuales contienen información sobre el año de publicación y el país de origen respectivamente. Para el criterio EC15, la distribución de los estudios por año de publicación se evidencia en la Figura 4.5, en donde observamos que hubo un crecimiento a partir del año 2016, llegando a un pico en el año 2019, pero con una caída en el año 2020. Este último es posiblemente debido a que la situación sanitaria que comenzó en el 2020 centró los esfuerzos de los investigadores en otras áreas diferentes a los desafíos del aprendizaje autodirigido. El número de estudios en el año 2021 es reducido debido a que la búsqueda de artículos fue ejecutada en marzo de ese año.

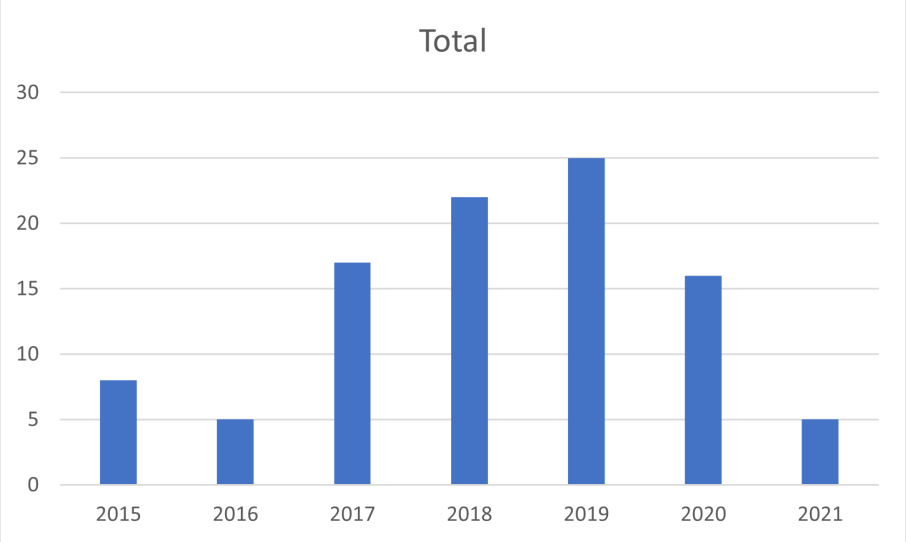

Figura 4.5: Estudios por año

En la Figura 4.6 podemos observar que el comportamiento de los estudios que tratan los desafíos del aprendizaje con minería de datos tiene un crecimiento a partir del 2016 llegando a su pico máximo en el 2019 y comenzando a reducirse en el 2020. Ahora bien, los estudios que tratan los desafíos del aprendizaje con minería de procesos aparecen recién en el año 2018 y no han tenido un aumento sustancial de casos con el pasar de los años, manteniendo pocos estudios por año hasta la fecha de la ejecución de las búsquedas para esta revisión.

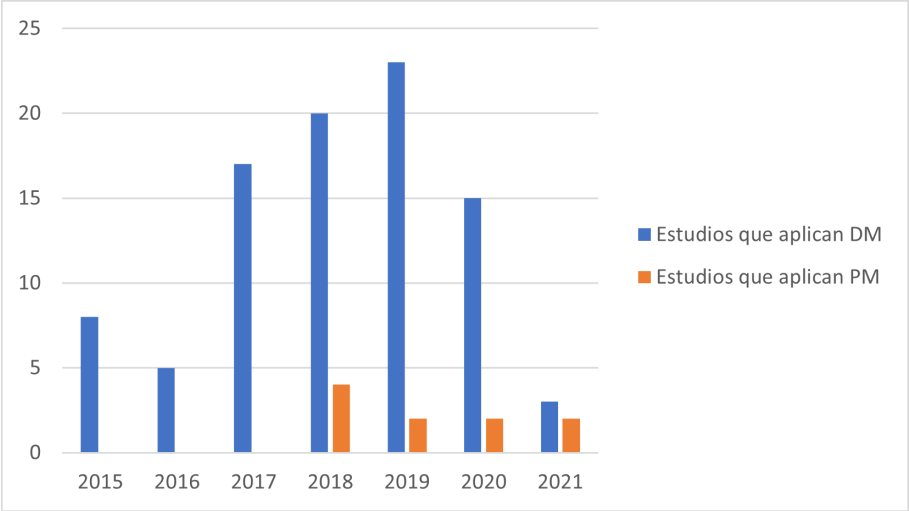

Figura 4.6: Estudios por año y por ciencia de la computación

Para el criterio EC16, en la distribución de estudios por país se observa que China es el principal proveedor de estas investigaciones con 40 estudios, seguido por USA con 12 estudios. Desde este punto de vista, los países de la Unión Europea aparecen mínimamente destacando los Países Bajos y España, mientras que, en cuanto a Latinoamérica el país que destaca es Chile.

Finalmente, se identificaron 68 fuentes de información entre revistas y conferencias. De estas 68 fuentes, 50 contienen solamente un artículo de los analizados, 12 contienen dos artículos y 4 contienen 3 artículos. Las fuentes con mayor cantidad de artículos son "Learning Analytics and Knowledge (LAK)" y "ACM Turning Conference (ACM TURC)". En la Tabla 4.6 se detalla las fuentes utilizadas en este estudio con más publicaciones.

Revista o conferencia	# de artículos
Learning Analytics and Knowledge	8
ACM Turning Conference	4
ACM Conference on User Modeling, Adaptation and Personalization	3
Intelligent Data Engineering and Automated Learning	3
Database Systems for Advanced Applications	3
European Conference on Technology Enhanced Learning	3

Tabla 4.6: Artículos por fuente

Análisis de contenido: En esta sección se analizan los resultados obtenidos para cada una de las sub-preguntas de investigación basándose en los criterios de extracción definidos en la tabla 3.2.

Primero, para la pregunta de investigación 1 (RQ1): ¿Qué técnicas de minería de datos o minería de procesos son utilizadas en los cursos MOOC? Se identificó ocho criterios de extracción:

- **EC01. Ciencia de la computación.** El estudio se enfocó en dos ciencias de la computación: La minería de datos y la minería de procesos. Basándose en esto se determinó que 91 estudios utilizaron minería de datos para analizar los desafíos del aprendizaje autodirigido, mientras que, 10 estudios utilizaron minería de procesos. De los 98 estudios analizados, tres de ellos utilizaron tanto minería de datos como minería de procesos. Por ejemplo, en el estudio realizado por Matcha et al. [24] se

analizó la autorregulación del aprendizaje utilizando técnicas de minería de procesos, apoyándose en técnicas de regresión logística para determinar la existencia de una relación entre la personalidad del estudiante y la estrategia de aprendizaje. Por otra parte, el estudio realizado por Maldonado [25] se centra de igual forma en determinar las estrategias de autorregulación de estudiantes en cursos MOOC, pero apoyándose en técnicas de clusterización clasificando a los estudiantes que muestran patrones de comportamiento similares. Finalmente, en otro estudio de Maldonado [26] se aplican técnicas de regresión lineal para determinar grupos de estudiantes con base en su comportamiento, siendo este comportamiento modelado mediante la aplicación de técnicas de minería de procesos. Se puede observar que los estudios que combinan la minería de datos con la minería de procesos aplican ambas ciencias como complemento una de la otra, en donde la minería de datos permite identificar de mejor manera grupos de datos para modelar procesos independientes mediante minería de procesos.

- **EC02. Técnicas de minería de datos.** En este estudio se identificó seis técnicas de minería de datos: Clasificación, Clusterización, Regresión, Reglas de asociación, Redes neuronales, Máquinas de vectores de soporte. En la Figura 4.7 sé puede observar que la técnica que encabeza los estudios es la clasificación, seguido por la regresión y la clusterización. Esto, en rasgos generales, si se analiza la distribución de las técnicas basándose en el desafío de aprendizaje abordado, la clasificación encabeza los estudios enfocados en el abandono como se pudo observar anteriormente en la figura 4.2. Por otro lado, la regresión es la técnica más utilizada para analizar la participación. La atención o interés se distribuye en partes iguales entre la clasificación y la regresión, mientras que los estudios que analizan la autorregulación utilizan en su mayoría técnicas de regresión. Finalmente, existe un único estudio que se centra en la procrastinación. Este es el estudio realizado por Li et al [27] y utiliza técnicas de clusterización para su propósito. Ahora bien, su estudio no se centra únicamente en la procrastinación, sino que también se enfoca en la participación.

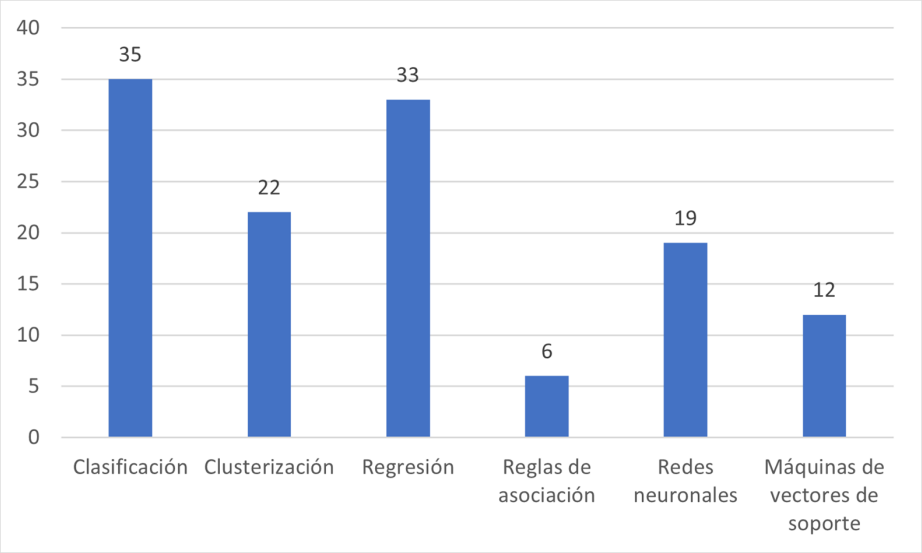

Figura 4.7: Estudios por técnica de minería de datos

- **EC03. Técnica de minería de procesos.** Para el análisis de este criterio de extracción se identificaron tres posibles técnicas de minería de procesos: Descubrimiento de procesos, Conformidad de procesos y Mejora de procesos. De los 98 estudios analizados, solamente 10 utilizan técnicas de minería de procesos, todos ellos enfocados

en el descubrimiento de procesos. De estos, 6 se enfocan en la autorregulación del aprendizaje y el recorrido realizado por los estudiantes basándose en ese desafío, mientras que los desafíos participación y el abandono son analizados en 2 estudios cada uno. No se identificaron estudios que analicen los desafíos de procrastinación y la atención o interés mediante estas técnicas. No se identificaron estudios que utilicen la conformidad de procesos o la mejora de procesos.

- **EC04. Algoritmo de minería de procesos.** El algoritmo de minería de procesos es un criterio abierto el cual, durante el proceso de lectura de los artículos identificó 10 algoritmos diferentes, en donde cada uno de los 10 estudios que utilizan técnicas de minería de procesos aportan por lo menos un algoritmo al listado. Destaca que dos estudios que analizan la autorregulación del aprendizaje utilizan el algoritmo Fuzzy Miner Disco, el cual es una adaptación específica del algoritmo Fuzzy Miner optimizado para la herramienta Disco Fluxicon [28]. Los algoritmos identificados fueron: Cspace, Secpi, PM2 method, Inductive miner, Heuristic algorithm, Fuzzy miner, EM algorithm, Epistemic network analysis, Fuzzy miner disco y Celonis algorithm.
- **EC05, EC06 y EC07.** Los criterios de extracción EC05 y EC06 son criterios relacionados con la técnica de minería de procesos de conformidad de procesos, al no haberse identificado estudios que utilicen esta técnica estos criterios no se utilizaron en ningún artículo. Por otra parte, el criterio EC07 es un criterio relacionado con la técnica de mejora de procesos, que al igual que la conformidad de procesos no fue utilizada, por lo que sus criterios relacionados no fueron utilizados.
- **EC08. Herramientas adicionales.** El criterio de herramientas adicionales es un criterio abierto, que surgió de la necesidad de identificar qué clase de herramientas utilizan los investigadores para brindar apoyo a la minería de datos y/o minería de procesos en sus estudios. Se identificaron siete herramientas y/o tecnologías adicionales. En estos destaca que los cuestionarios tienden a utilizarse en estudios que analizan la autorregulación del aprendizaje. Dichos cuestionarios se toman como la base sobre la cual predecir el nivel de autorregulación. Por otra parte, el uso de estadística destaca en estudios que analizan la participación y la autorregulación, principalmente en el preprocesamiento de datos del alumno. Esta relación entre las herramientas y los desafíos del aprendizaje se aprecia en la Figura 4.4.

Para la pregunta de investigación 2 (RQ2): ¿En qué desafíos del aprendizaje autodirigido se han aplicado técnicas de minería de datos o minería de procesos? Se identificó dos criterios de extracción:

- **EC09. Desafíos del proceso de aprendizaje autodirigido en entornos online.** Para este criterio se identificó cinco desafíos del proceso de aprendizaje a los que se enfrentan estudiantes y tutores en el contexto de cursos MOOC. Estos son el abandono, la procrastinación, el interés o motivación, la participación y la autorregulación del aprendizaje. La participación es el desafío con más estudios en la muestra, abarcando 49 estudios. En estos, los datos mayoritariamente analizados fueron los registros de actividad en foros de discusión del curso. Seguido a la participación tenemos el abandono, con 39 estudios que lo tratan. Para ello se analizan datos de interacción con la plataforma y las variaciones que estos presentan antes de que el estudiante deje de acceder al curso. Luego, la motivación o interés se mide en 17 estudios trabajando con registros de actividad con el material del curso. La autorregulación del aprendizaje se trata en 10 estudios, igualmente trabajando con registros de actividad. Finalmente, la procrastinación se trata únicamente en un artículo.

- **EC10. Tipo de datos analizados.** Se identificaron tres categorías de datos utilizados en los artículos analizados: La información demográfica, los registros de actividad o registros de eventos y la información de evaluación. Los registros de actividad, ya sea en el material educativo o en foros de discusión, abarcan 90 de los 98 artículos analizados. La información demográfica e información de evaluación son utilizados en 22 y 14 artículos respectivamente. Gran parte de estos complementan a los registros de navegación.

Finalmente, para la pregunta de investigación 3 (RQ3): ¿Cómo se aborda la investigación de los desafíos del aprendizaje autodirigido en cursos MOOC mediante la aplicación de técnicas de minería de datos o minería de procesos? Se identificaron seis criterios de extracción, donde dos de ellos fueron ya detallados en el análisis de metadatos. Los cuatro restantes fueron:

- **EC11. Fase.** Para este estudio se consideró que el proceso de investigación consta de cuatro fases: Análisis, Diseño, Implementación y Testeo. Para los estudios analizados gran parte de ellos pasan por la fase de análisis, siendo esta fase mayoritaria presente en 91 de los 98 artículos. Muy por detrás se encuentran la fase implementación con 6 artículos y las fases de diseño y testeo con tres en cada una. Hay que tener en cuenta que en el caso de los estudios que pasan por la fase de diseño y/o implementación se está trabajando con versiones híbridas de algoritmos de minería de datos existentes, por lo que era necesario verificar su correcto funcionamiento.
- **EC12. Tipo de validación.** Se identificó cinco tipos de validación para los estudios: Prueba de conceptos, Encuesta, Experimento, Prototipo y Caso de estudio. En estos, los investigadores mayoritariamente validan su estudio mediante experimentos [1] con 55 artículos. El segundo tipo de validación más utilizado son los casos de estudio [2] con 47 artículos empleando esta forma de validación. Finalmente, la validación mediante prototipos aparece mínimamente. Las pruebas de conceptos y encuestas no son utilizados como tipos de validación en los estudios analizados.
- **EC13. Alcance del enfoque.** Al tratarse de una tecnología educativa, todos los artículos que forman parte de esta muestra tienen un enfoque académico.
- **EC14. Metodología.** Todos los artículos son nuevos. Ninguno de los pertenecientes a esta muestra es una extensión de estudios anteriores. Ahora bien, algunos de ellos son rectificaciones de información básica de los autores o los datos analizados, más no una modificación del resultado. Para estos casos se incluyó en la muestra únicamente la versión corregida del artículo.

4.2 Crítica a la revisión sistemática

En este capítulo, se realizó una revisión sistemática de literatura para identificar qué desafíos del aprendizaje autodirigido en el contexto de cursos MOOC han sido objeto de estudio mediante la aplicación de técnicas de minería de datos o minería de procesos. Para ello se usó una versión adaptada de la metodología de Barbara Kitchenham para revisiones sistemáticas. De este modo, se plantearon tres preguntas de investigación vinculadas a un conjunto de criterios de extracción de información de los artículos analizados.

[1] Llámese experimento a estudios que utilizan datos procedentes de otras fuentes, ya sea repositorios de datos u otros estudios

[2] Llámese caso de estudio a artículos que utilizan datos propios o específicos de un escenario bien definido en concreto

Los resultados muestran que la minería de procesos no es una tecnología que se utiliza comúnmente en el ámbito de la educación en línea en el contexto de cursos MOOC. Para abordar los desafíos del aprendizaje, gran parte de los estudios se han centrado en analizar el comportamiento y el abandono, y la posible existencia de una relación entre estos, mediante la aplicación de técnicas de minería de datos. Otros aspectos, como la motivación o la procrastinación, han sido rezagados a contados estudios.

Predecir el abandono de los estudiantes es importante para cualquier ámbito educativo, ya sea presencial o en línea, pero identificar patrones de comportamiento que podrían desembocar en el éxito o fracaso de un alumno en el MOOC es un gran aporte a esta área de estudio. Incluso permitiría determinar si las variaciones de comportamiento o la asimilación de formas inadecuadas de llevar el material educativo afectan al resultado final del mismo.

Esta revisión sistemática pretende identificar los vacíos en la investigación de los desafíos del aprendizaje autodirigido con el fin de servir como base para determinar la temática de este trabajo fin de máster. Por lo tanto, este trabajo analizará, mediante la aplicación de técnicas de minería de procesos, uno de los desafíos menos abarcados por los estudios analizados: la procrastinación. Con esto se persigue identificar si este comportamiento en los cursos MOOC tiene relación alguna con el éxito o el fracaso de sus participantes.

CAPÍTULO 5
Solución propuesta

Como se puede observar en los resultados obtenidos en el capítulo anterior, a partir del año 2016 se manifestó en la comunidad de investigadores un aumento considerable en el interés de la aplicación de técnicas de minería de datos sobre los registros generados en cursos MOOC. Gran parte de estos estudios están orientados principalmente a predecir o determinar un punto de abandono del curso.

Por otra parte, otra ciencia de la computación surgió en el año 2018, pero a menor medida, siendo esta la minería de procesos y cuyos estudios se han centrado principalmente en analizar el recorrido de grupos definidos de estudiantes basándose en diferentes características como, por ejemplo, su nivel de autorregulación.

De los cinco desafíos que afronta el aprendizaje autodirigido en cursos en línea, uno de los menos analizados es la procrastinación. La solución propuesta en este TFM es la aplicación de técnicas de minería de procesos en el descubrimiento de patrones de procrastinación en el comportamiento de usuarios en un curso MOOC utilizando, también, minería de datos como punto de apoyo en el desarrollo del caso de estudio.

5.1 Diseño de la solución

Para la implementación de la solución propuesta en este TFM se utiliza de forma adaptada la metodología PM^2 de minería de procesos propuesta por Maikel L. van Eck [29]. Esta versión se encuentra descrita en la figura 5.1.

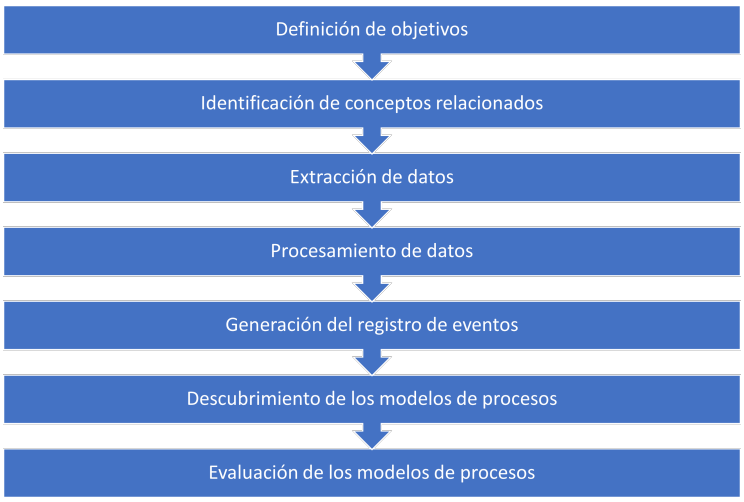

Figura 5.1: Metodología adaptada de PM^2

5.2 Diseño detallado

En este apartado, se detalla a grandes rasgos las actividades a realizar en cada una de las etapas de la metodología PM^2 adaptada para la ejecución de este caso de estudio.

5.2.1. Definición de objetivos

Inicialmente, se planteó un conjunto de objetivos en el capítulo introductorio, los cuales es necesario cumplir para llegar a la meta principal de este estudio. Por ello, en esta etapa de la metodología es necesario establecer las preguntas de investigación a responder mediante el caso de estudio y la ejecución de los diferentes experimentos que competen a éste.

5.2.2. Identificación de conceptos relacionados

En esta etapa de la metodología se requiere la identificación de las diferentes bases teóricas del comportamiento de participantes de cursos en línea. Su utilidad radica en tener una comprensión más clara del comportamiento de los estudiantes de los cursos MOOC. De este modo, se podrá realizar una identificación adecuada de los modelos de procesos basándose en los datos a analizar y que los resultados se adapten a estas bases teóricas.

5.2.3. Extracción de datos

Para continuar con la metodología planteada, en esta etapa se procede a obtener los datos necesarios de la plataforma de distribución de cursos MOOC que amerita el caso de estudio. La plataforma a utilizar, en este caso, es OpenEDX, manejada por la Universidad Politécnica de Valencia.

5.2.4. Procesamiento de datos

Antes de proceder a generar los registros de eventos es necesario procesar los datos obtenidos en su estado inicial. De este modo, se adaptan los conceptos relacionados identificados en la etapa dos. Además, mediante la aplicación de técnicas de minería de datos, se definen los diferentes grupos de estudiantes a tratar.

5.2.5. Generación del registro de eventos

Al tener los datos ya procesados, la siguiente etapa de la metodología consiste en la generación del registro de eventos limpios de los diferentes grupos de participantes del curso identificados mediante la aplicación de técnicas de minería de datos. A su vez, estos grupos deben de cumplir con las bases teóricas de comportamiento identificadas anteriormente y que guardan relación con los objetivos del estudio.

5.2.6. Descubrimiento de los modelos de procesos

Mediante la aplicación de técnicas y algoritmos de minería de procesos se procede con la generación de los modelos de procesos a ser analizados. Dichos modelos serán la base principal para dar respuesta a las preguntas de investigación planteadas.

5.2.7. Evaluación de los modelos obtenidos

Una vez que se tiene los diferentes modelos de procesos generados, la última etapa de la metodología consiste en analizar los resultados obtenidos. A partir de estos, basándose en el comportamiento que los modelos de procesos reflejan, se da respuesta a las preguntas de investigación planteadas.

5.3 Tecnología utilizada

Para la correcta ejecución de la solución propuesta se requiere el uso de diferentes herramientas tecnológicas que se adapten a las necesidades del estudio. En este caso en particular, se define la necesidad de utilizar principalmente dos herramientas de software: R y Disco Fluxicon.

5.3.1. R

R es un entorno integrado de software de código abierto que se utiliza principalmente en la manipulación, análisis, cálculo, visualización y gestión de datos. La capacidad adaptativa de R le permite ser un entorno flexible a las diferentes circunstancias que requieran de su uso [30].

Figura 5.2: Logotipo del lenguaje de programación R

5.3.2. Disco Fluxicon

Disco Fluxicon es una herramienta comercial de minería de procesos. Esta se caracteriza, principalmente, por ser independiente y liviana. Presenta flexibilidad en el formato de sus archivos de entrada, aunque su soporte nativo se centra en el formato CSV. La herramienta DISCO fue construida con el fin de garantizar usabilidad, fidelidad y un buen rendimiento por lo que su curva de aprendizaje garantiza que la aplicación de las técnicas de minería de procesos que dispone sean fáciles y rápidas para el usuario. [31]

Figura 5.3: Logotipo de Disco FLUXICON

CAPÍTULO 6
Caso de estudio

La Universidad Politécnica de Valencia oferta anualmente dos ediciones del curso "Basic Spanish 1: Getting Started". Éste está dirigido a personas angloparlantes de todo el mundo. Para este caso de estudio, se utilizó la información de los estudiantes matriculados en la primera temporada del año 2020. En el presente capítulo se detalla el caso de estudio donde, mediante la aplicación de técnicas de minería de procesos, se intenta identificar variaciones en el comportamiento de los estudiantes clasificados por grupos con el fin de determinar la existencia de una relación entre la procrastinación y el éxito y/o fracaso de un estudiante en un curso en línea.

6.1 Definición del caso de estudio

6.1.1. Objetivo

El objetivo de este caso de estudio es identificar los patrones de comportamiento relacionados con la procrastinación en los estudiantes en el curso MOOC basándose en el resultado obtenido por los mismos al finalizar el curso. Para cumplir con este objetivo, se realizó el análisis sobre los datos del registro de eventos de los estudiantes resultado de su navegación por el curso.

6.1.2. Contexto

El curso MOOC utilizado fue "Basic Spanish 1: Getting Started" ofertado por la Universidad Politécnica de Valencia en su plataforma OpenEDX. El curso tuvo una duración de 7 semanas, sin contar la semana de introducción y las semanas dedicadas a la evaluación intermedia y final. Se registraron un total de 136375 estudiantes, de los cuales 96815 estudiantes registraron actividad. El curso se ofertó de forma abierta a nivel global, principalmente, para personas que provenían de países angloparlantes.

6.1.3. Curso MOOC

El curso se estructuró en 7 módulos de contenido, un módulo introductorio y dos módulos de evaluación. Cada módulo está conformado por un conjunto de lecciones, las cuales pueden ser: Lectura, Video, Audio, Pronunciación o Evaluación.

El curso cuenta con un total de 32 lecturas en formato HTML, 58 videos, 29 audios, 36 actividades de pronunciación y 89 actividades de evaluación distribuidos como se muestra en la tabla 5.1, en donde, cada asterisco (*) representa una lección. Como se

puede observar en la figura 6.1, el tipo de lecciones predominante en las identificadas son las actividades de evaluación, que representan el 36 % del contenido. Ahora bien, es necesario poner en consideración lo siguiente:

1. Las lecciones de lectura, video y actividades de evaluación fueron contabilizadas por elemento, es decir, que para las actividades de evaluación cada pregunta cuenta como un elemento independiente en el contador.

2. Las lecciones de audio y pronunciación se consideraron en conjunto. Esto es debido a que en los datos analizados en la interacción sobre estos elementos se presenta por el grupo al que pertenecen, más no de forma individual.

	Lectura	**Video**	**Audio**	**Pronunciación**	**Evaluación**
Semana 0. Course Welcome					
Welcome!	*	*			
Semana 1. Presentarse					
Presentación	*	**			
Saludar	*	*****	***	****	
El nombre	*	**		**	*
El alfabeto			*	*	*
Letras y sonidos	*	*******		*******	
Recapitulando	*				
Semana 2. De dónde eres?					
Presentación	*				
El verbo ser		*	*		***
La nacionalidad	*	*****	*	*	****
Los idiomas	*	***	*	**	*
Los numeros del 0 al 20		*			**
La hora			*		*
Recapitulando	*				
Semana 3. La familia					
Presentación	*				
Los parientes			**	*	*******
Los posesivos	*	***	*	*	*
El físico	*	******		**	*****
Recapitulando	*				
Semana 4. La ropa					
Presentación	*				
El vestuario	*	***	**	****	*
El artículo	*	**			****
Los colores			*	**	*****
Preguntas y respuestas	*				****
Recapitulando	*				
Mid-term exam					
Mid-term					*****
Semana 5. La casa					
Presentación	*				
Cómo es tu apartamento?			****	****	***
Objetos de una casa			*	*	*
Mobiliario		*	**	***	**
Está a la derecha			*	*	*

Grande y pequeño	*				**
La contracción		**			*
Recapitulando	*				
Semana 6. Dónde está?					
Presentación	*				
Estar	*	***			*
Hay unos cuadros		*	*		*
Esta aqui		*			***
Semanas y meses			**		**
Recapitulando	*				
Semana 7. En la calle					
Presentación	*				
Productos			***		**
Las tiendas			*		***
Dónde hay una farmacia?		****			
Viajar en tren	***				***
Los numeros		****			**
Practicando					**
Recapitulando	*				
Examen final					
Presentación	*				
Examen Final					******** ******
¡Hasta la próxima!	*				

Tabla 6.1: Estructura del curso

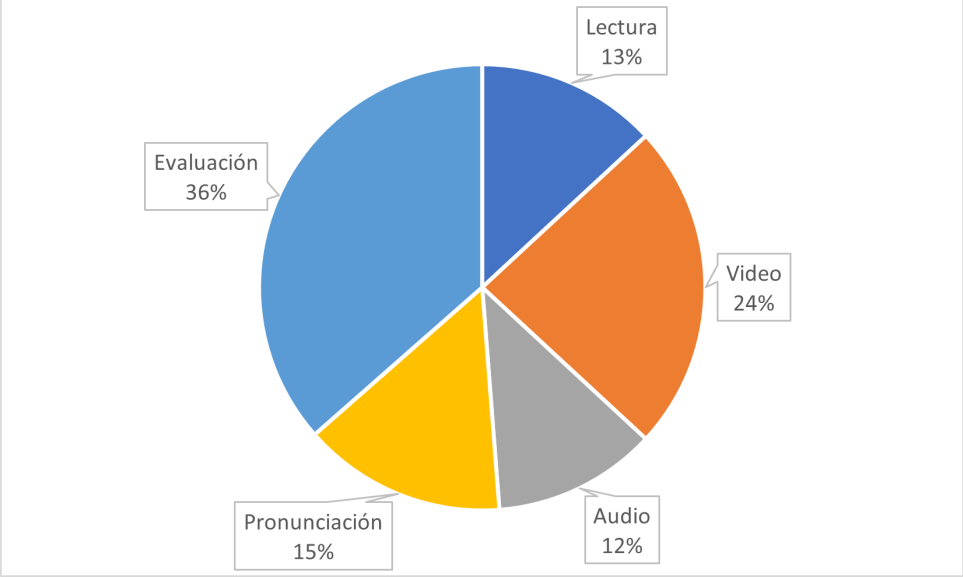

Figura 6.1: Distribución de contenido de lecciones del curso

6.1.4. Mediciones e instrumentos

De los estudiantes matriculados, 39560 no registran actividad sobre el curso. Por lo que, para este análisis, se consideró solamente a los estudiantes que presentaron acti-

vidad en el curso, siendo estos un total de 96815 participantes. En la figura 6.2 se puede observar el progreso de abandono de los estudiantes del curso. En la tabla 6.2 se especifica la cantidad de estudiantes del total que proporcionaron información demográfica detallada. En las tablas 6.3, 6.4 y 6.5 se encuentra la información demográfica de los estudiantes en lo que compete a género, rango de edad y nivel de preparación respectivamente.

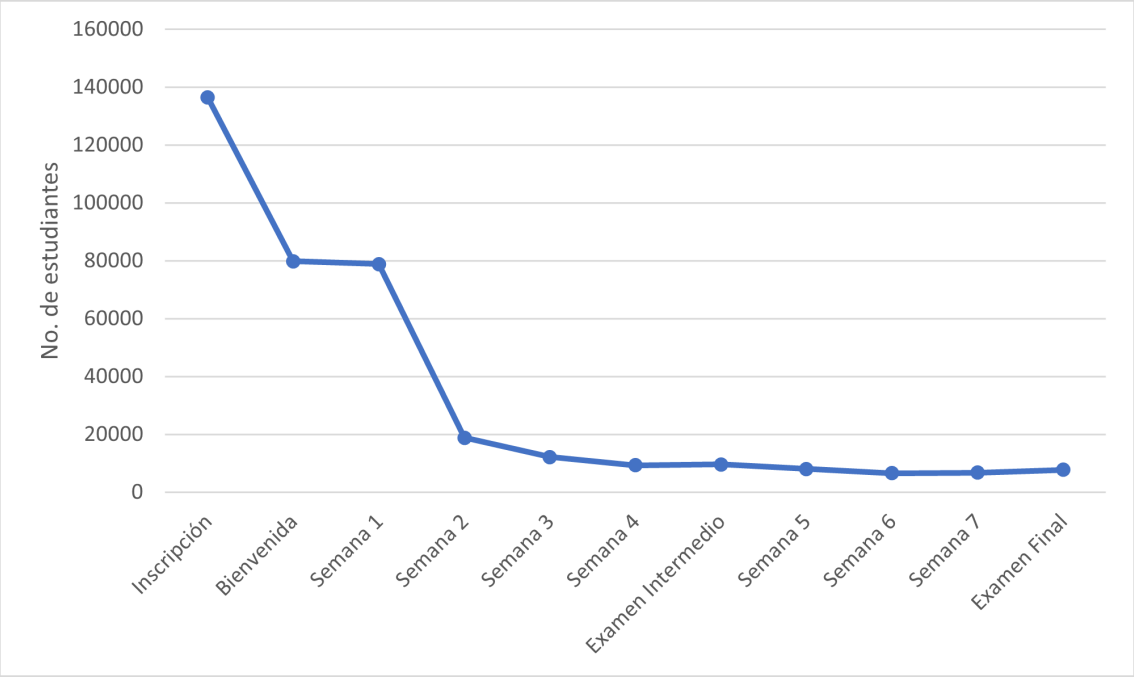

Figura 6.2: Flujo de estudiantes

Información	Responde	No responde
Género	58,20 %	41,80 %
Década	55,90 %	44,10 %
Preparación académica	62,39 %	37,61 %

Tabla 6.2: Participación en entrega de datos demográficos

Género	Cantidad	Porcentaje
Hombre	20038	35,02 %
Mujer	36699	64,15 %
Otro/Prefiere no decir	474	0,83 %

Tabla 6.3: Género

Década	Cantidad	Porcentaje
1920	4	0,01 %
1930	28	0,05 %
1940	343	0,57 %
1950	1279	2,12 %
1960	2687	4,45 %
1970	4525	7,50 %
1980	10292	17,05 %
1990	28299	46,89 %
2000	12760	21,14 %
2010	91	0,15 %
2020	41	0,07 %

Tabla 6.4: Rangos de edad

Nivel de preparación académica	Cantidad	Porcentaje
Associate degree	3041	5,91 %
Bachelor's degree	20038	38,93 %
Doctorate	1257	2,44 %
Doctorate in another field	17	0,03 %
Doctorate in science or engineering	16	0,03 %
Elementary/primary school	276	0,54 %
Junior secondary/junior high/middle school	1700	3,30 %
Master's or professional degree	10038	19,50 %
No Formal Education	149	0,29 %
Other Education	940	1,83 %
Secondary/high school	13994	27,19 %

Tabla 6.5: Nivel de preparación académica

6.2 Aplicación de la metodología PM^2

6.2.1. Objetivo y preguntas de investigación

Con base en el objetivo general planteado en el capítulo 1, el cual es "Explorar patrones de procrastinación en el comportamiento de los participantes de un curso MOOC, y determinar la relación que existe entre este comportamiento contraproducente y el éxito o fracaso de las personas inscritas en el curso" se consideraron las siguientes preguntas de investigación de interés para este estudio.

RQ1. ¿Existen variaciones en el comportamiento de los diferentes grupos de participantes en un curso MOOC?

RQ2. ¿Cuáles son las principales características del comportamiento de los diferentes grupos de participantes en un curso MOOC?

RQ3. ¿Existe relación entre la procrastinación y el éxito o fracaso de un participante en un curso MOOC?

6.2.2. Conceptos relacionados

Para una correcta interpretación de los resultados obtenidos en este estudio, es necesario que se presente al lector un conjunto de conceptos teóricos y bases de información necesarias. En esta subsección se exponen dichos conceptos y conocimientos requeridos.

Descripción de los datos

La plataforma de Open EDX almacena la información de los usuarios y los cursos en dos bases de datos:

1. **MySQL:** Es un sistema de gestión de bases de datos relaciones de código abierto y gratuito, con bajos costos de implementación. MySQL se caracteriza por ser uno de los sistemas gestores de bases de datos que más rápido ejecuta sus consultas de información [32].

2. **MongoDB:** Es una base de datos NoSQL de código abierto orientada a documentos. Las bases de datos NoSQL se caracterizan por no tener un esquema de tablas establecido para almacenar los datos. El tener un esquema no definido les permite a las bases de datos NoSQL tener una mejor escalabilidad, pero, con una menor coherencia de los datos [32].

La estructura de los datos utilizados en este estudio es descrita a continuación. Es necesario considerar que, debido a la flexibilidad de almacenamiento de MongoDB, las estructuras que provengan de esta base de datos no incluirán el tipo de datos para cada campo.

1. **Información del curso.**La información de los cursos se almacena en MongoDB [33]. Para este estudio se obtuvo esta información en un archivo CSV. El diccionario de datos de este archivo se describe en la tabla 5.6.

Columna	Descripción	Tipo
course_id	Identificador único del curso, el cual suele estar presente en las diferentes URL de su contenido	-
platform_id	Identificador de la plataforma que contiene al curso	-
title	Nombre público del curso, el cual se muestra en pantalla para los potenciales usuarios	-
start_date	Fecha de inicio del curso	-
weeks	Número total de semanas que conforman el curso, en este número se incluyen la semana de introducción y las diferentes semanas de evaluación	-
lang	Lenguaje en el que se oferta el curso, este campo no representa el lenguaje que el curso imparte (en este caso español), sino el idioma del material básico del mismo, es decir, indicaciones generales, etc.	-
pass_grade	Nota mínima que requieren los participantes que pagaron por el certificado del curso para obtenerlo	-

short_desc	Descripción corta en formato HTML que permite al usuario o potencial usuario del curso conocer a breves rasgos el contenido de este.	-
url	Dirección URL completa del curso	-
end_date	Fecha limite a la que los estudiantes que se matricularon en esta edición del curso pueden obtener el certificado o pagar por este.	-

Tabla 6.6: Diccionario de datos para el archivo de información del curso

2. **Información demográfica.** La información demográfica de los usuarios del curso se almacena en MySQL [33] en la tabla "auth_userprofile" [34]. El diccionario de estos datos se detalla en la tabla 6.7.

Columna	Descripción	Tipo
country	Almacena el código de país del estudiante en un código del estándar "GENC 2-letter code" basado en la "ISO 3166-1 alpha-2". Actualmente el registro de país es obligatorio, antes se guardaba en blanco cuando el usuario no registraba esta información.	varchar(2)
level_of_education	Información de la formación académica del usuario, a partir de un conjunto de opciones detallados en la tabla 5.8.	varchar(6)
gender	Información de género del usuario, a partir de un conjunto de opciones detalladas en la tabla 6.9	varchar(6)
year_of_birth	Año de nacimiento del usuario, recopilado durante el registro, con valor NULL para quienes no respondieron	int(11)
user_id_encrypt	Identificador encriptado único del usuario	int(11)

Tabla 6.7: Diccionario de datos para el archivo de información demográfica de los participantes del curso

Código	Descripción
p	Doctorate
m	Master's or professional degree
b	Bachelor's degree
a	Associate degree
hs	Secondary/high school
jhs	Junior secondary/junior high/-middle school

el	Elementary/primary school
none	No Formal Education
other	Other Education
(blank)	El usuario no especificó el nivel de educación
p_se	Doctorate in science or engineering (ya no se usa)
p_oth	Doctorate in another field (ya no se usa)
NULL	Para un alumno que no respondió o que se registró antes de que esta información sea obligatoria

Tabla 6.8: Opciones para registro de formación académica

Código	Descripción
f	Female
m	Male
o	Other/Prefer Not to Say
(blank)	User did not specify a gender
NULL	Para un alumno que no respondió o que se registró antes de que esta información sea obligatoria

Tabla 6.9: Opciones para registro de género

3. **Información de inscripción.** La información de inscripción de los usuarios se almacena en MySQL [33] en la tabla "student_courseenrollment" [34]. El diccionario de datos de esta tabla se encuentra en la tabla 5.10.

Columna	Descripción	Tipo
course_id	Identificador único del curso en el cual el usuario se ha registrado	varchar(255)
created	Almacena la fecha y hora de registro del usuario en el curso en formato UTC	datetime
is_active	Indicador booleano que indica si la inscripción del usuario en el curso se encuentra activa, una inscripción inactiva se representa con el valor 0 (Falso). El cambio de este indicador no implica la eliminación de datos del usuario en el curso, por lo que si el alumno vuelve al curso sus datos anteriores se mantienen.	tinyint(1)

mod	Cadena de caracteres que indica el tipo de inscripción de los usuarios del curso, para este caso de estudio se identificó dos categorías: Audit (inscripción sin certificado) y Verified (inscripción con certificado)	varchar(100)
user_id_encrypt	Identificador encriptado único del usuario	int(11)

Tabla 6.10: Diccionario de datos para el archivo de información de registro de usuarios en el curso

4. **Información de finalización** La información de finalización del curso de los usuarios se almacena en MySQL [33] en la tabla "grades_persistentcoursegrade" [34]. Los datos contenidos por esta tabla se relacionan únicamente con los usuarios que finalizaron el curso. Su diccionario de datos se encuentra detallado en la tabla 5.11.

Columna	Descripción	Tipo
created	Timestamp en que se calculó por primera vez la calificación del usuario para el curso	DateTime
modified	Timestamp en que se actualizó por última vez la calificación del usuario para el curso	DateTime
course_id	Identificador único del curso en el cual el usuario obtuvo la calificación registrada	CourseKey
course_edited_timestamp	Último Timestamp en que se calculó la calificación, se utiliza únicamente para depurar	DateTime
grading_policy_hash	Política criptográfica de tipo SHA-1 que permite a la plataforma detectar y actualizar las calificaciones de los usuarios cada que el curso cambia su política de evaluación	String (255)
percent_grade	La calificación del alumno calculada por el curso como un porcentaje decimal, basándose en la política de calificación.	Float
letter_grade	La calificación del alumno calculada por el curso como un valor de una cadena de caracteres basándose en la política de calificación.	String (255)

passed_timestamp	Timestamp en que el usuario aprobó el curso por primera vez, si este valor se encuentra vacío el usuario nunca aprobó el curso. Si este valor se encuentra con información, pero la columna letter_grade se encuentra vacía, quiere decir que el alumno pasó de un estado aprobado ha reprobado.	DateTime
user_id_encrypt	Identificador encriptado único del usuario	Integer

Tabla 6.11: Diccionario de datos para el archivo de información de finalización del curso

5. **Registros de navegación** La información que compete a los registros de navegación de los estudiantes se guarda en dos instancias: MongoDB y MySQL [33]. Para este caso de estudio, se tuvo acceso únicamente a la información contenida en MySQL a partir de la tabla "courseware_studentmodule" [34]. Su diccionario de datos se describe en la tabla 5.12.

Columna	Descripción	Tipo
module_type	Todos los cursos están conformados con un conjunto de módulos de diferentes niveles en un orden específico, este campo identifica el tipo de módulo que fue visitado por el usuario, en una subsección posterior se detalla de mejor manera los tipos de módulos que conforman el curso que compete a este caso de estudio	varchar(32)
module_id	Identificador único del módulo que fue visitado por el usuario	varchar(255)
course_id	Identificador único del curso al que pertenece el módulo visitado por el usuario	varchar(255)
grade	Valor de punto flotante que indica la calificación no ponderada del usuario en el módulo en el caso de que este sea de tipo "problem"	double
max_grade	Valor de punto flotante que indica la calificación no ponderada máxima posible a obtener en el módulo en el caso de que este sea de tipo "problem"	double

created	Fecha y hora en que se creó el registro de navegación, se genera cuando el usuario visita un módulo por primera vez, comúnmente si ese módulo tiene uno o más módulos hijos se crea un registro por cada hijo con la misma fecha y hora de su módulo padre.	datetime
modified	Fecha y hora en que se actualizó el registro de navegación, comúnmente este valor comienza siendo el mismo que el de la fecha y hora de creación, pero cambia en cuanto el usuario interactúa con el módulo	datetime
user_id_encrypt	Identificador encriptado único del usuario	int(11)

Tabla 6.12: Diccionario de datos para el archivo de registros de navegación de los usuarios

Estructura del curso

Todo curso de la plataforma EDX está conformado por un conjunto de componentes llamados módulos, los cuales pueden ser de una diversa cantidad de categorías. Dependiendo de su tipo, estos contienen otros módulos a los cuales llamaremos módulos hijos. Para este caso específico, se han identificado 7 tipos de módulo utilizados, los cuales son:

1. **Course.** El módulo de tipo "course" es el módulo padre de todo curso de la plataforma Open EDX, sus módulos hijo directos son los módulos de tipo "chapter".

2. **Chapter.** El módulo de tipo "chapter" es el módulo que referencia al contenido semanal del curso de manera general, sus módulos hijo directos son los módulos de tipo "sequential".

3. **Sequential.** El módulo de tipo "sequential" es el módulo que contiene a los subcapítulos de cada módulo de tipo "chapter". Este módulo gestiona de manera general el contenido del curso, mientras que, sus módulos hijo gestionan el contenido de forma específica. Sus módulos hijo directos son los módulos de tipo "vertical".

4. **Vertical.** Los módulos de tipo "vertical" son aquellos que gestionan de forma general las lecciones del curso. Cada uno de estos módulos pueden contener una o varias lecciones, las cuales también son módulos, pero de otras categorías. Sus módulos hijo directos pueden ser de tipo: "video", "problem" o "drag-and-drop-v2".

5. **Video.** Los módulos de tipo "video" son aquellos que contienen material audiovisual ya sea nativo, es decir, contenido dentro del mismo servidor de la plataforma, o externo en otros servicios como YouTube.

6. **Problem.** Los módulos de tipo "problem" son aquellos que contienen actividades de evaluación de diversos tipos. Estas actividades pueden ser evaluadas o no evaluadas dependiendo de la configuración del curso.

7. **Drag-and-drop-v2.** Los módulos de tipo "drag-and-drop-v2" son un tipo especial de módulo de problema que se caracteriza por contener ejercicios especiales en los que se arrastra y suelta diferentes opciones en un espacio específico. Al igual que los módulos de tipo "problem", pueden ser o no evaluados dependiendo de la configuración.

En la figura 6.3 se puede observar la estructura jerárquica de los diferentes módulos dentro del curso, mientras que, en las figuras 6.4 y 6.5 se puede observar la estructuración del contenido de forma visual.

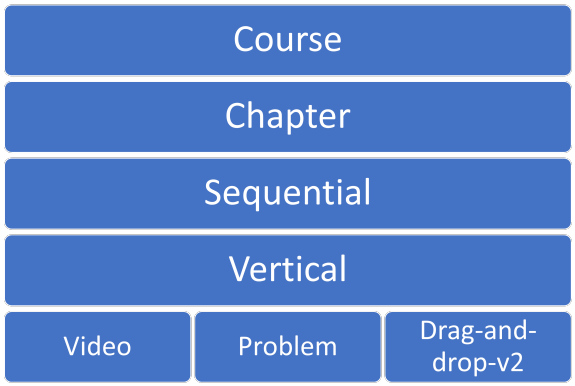

Figura 6.3: Estructura jerárquica del curso

Figura 6.4: Estructura del curso - parte 1

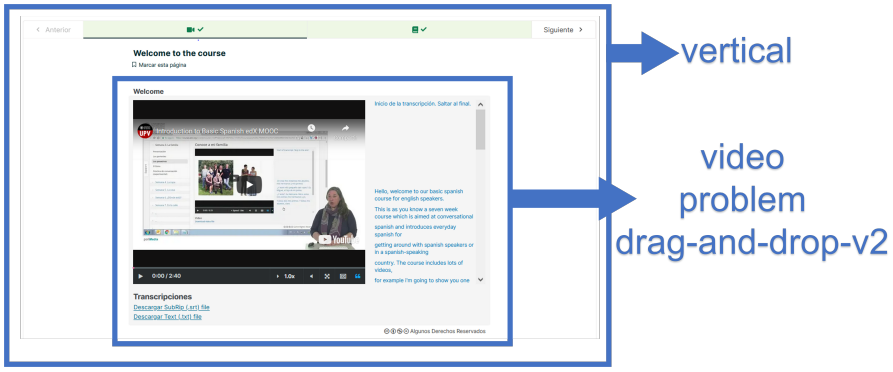

Figura 6.5: Estructura del curso - parte 2

Sesión de estudio

Una sesión de estudios puede definirse como un conjunto de interacciones con las lecciones de la plataforma ordenadas cronológicamente, las cuales ocurren dentro de un periodo determinado de tiempo [35].

Los investigadores han identificado y/o ideado diferentes formas de definir una sesión de estudio dentro de un curso MOOC. Por ejemplo, el estudio realizado por Yu et al. [36] lo realiza de una forma fácil, puesto que la plataforma de la cual se origina su caso de estudio, la plataforma OpenEdu, proporciona un identificador único de sesión por cada una de ellas. Esto permite a los investigadores omitir el proceso de identificación de una sesión y que sus estudios tengan una mayor precisión con respecto a los datos analizados. Ahora bien, no todos los estudios tienen esta facilidad.

Los datos analizados en este estudio no cuentan con una diferenciación de sesiones entre la actividad de los usuarios, por lo que es necesario plantear una estrategia de identificación de sesiones. Una práctica común de otros estudios es la definición de un tiempo determinado de inactividad, denominado umbral de inactividad, el cual, al cumplirse, indica el inicio de una nueva sesión de estudios. Existen varios ejemplos en los que se define este umbral de actividad, como es el caso de los estudios realizados por Ren et al. [37] y Vitiello et al. [35] quienes definen umbrales de inactividad de una hora y de 30 minutos respectivamente, pero no especifican el porqué de esta decisión.

Un tercer estudio realizado por Barba et al. [38] también toma en cuenta un umbral de inactividad de 30 minutos, pero, a diferencia de los dos anteriores mencionados, este estudio hace hincapié en especificar el porqué de dicha decisión. De este modo, para el curso analizado en su caso de estudio existían videos de hasta 23 minutos. Eso, sumado a un posible tiempo de reflexión del material analizado, da como resultado el tiempo del umbral de inactividad seleccionado.

El caso de estudio tratado en este documento, además de tener videos, también tiene material auditivo y ejercicios de pronunciación. Por ello, para definir el umbral de inactividad a utilizar, es necesario considerar los tiempos esperados que contemplan estas lecciones.

Los tiempos de duración máximo y mínimo de un video de los datos analizados se muestran en la figura 6.6. Donde se puede observar que, el video más largo tiene cinco minutos con cinco segundos de duración, mientras que el video más corto tiene una duración de cinco segundos.

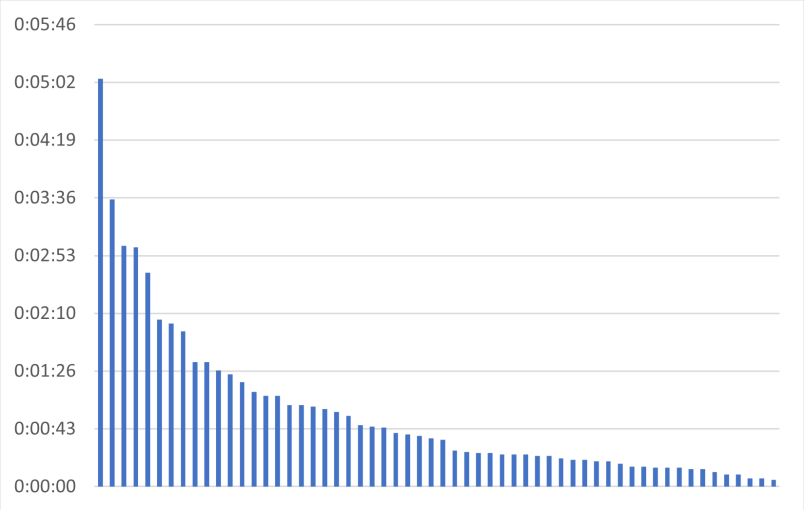

Figura 6.6: Duración de videos

Por otra parte, el tiempo de duración del material de audio es presentado en la figura 6.7. En esta figura se puede observar que el material de audio tiene una duración mucho menor al material de video. Donde, el mayor tiempo registrado, es de un minuto con 31 segundos, mientras que el menor es de tres segundos.

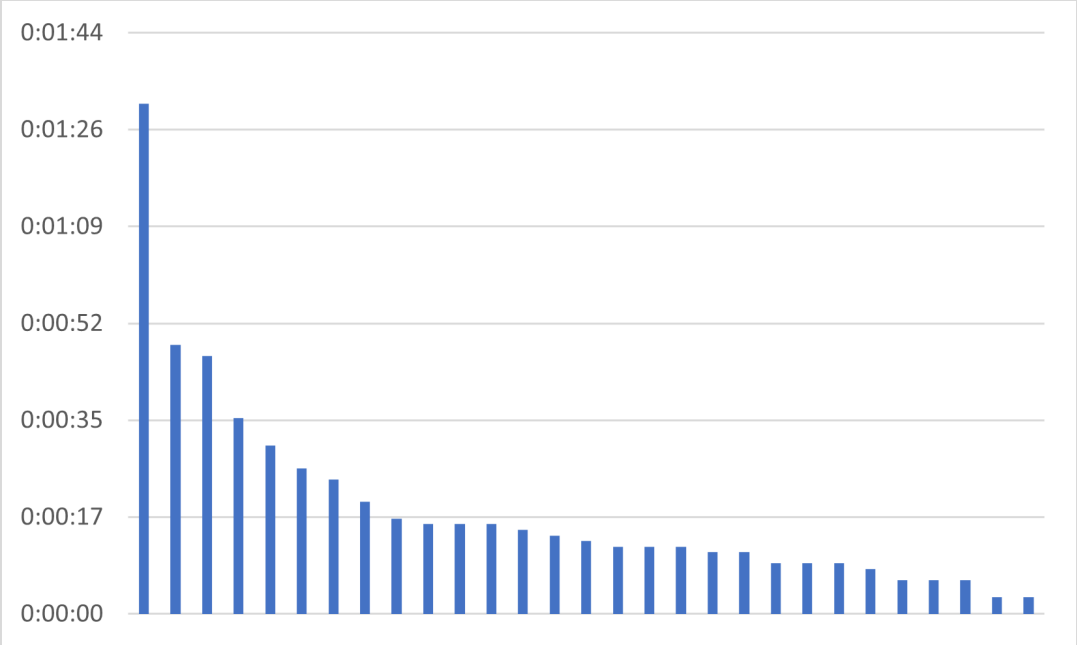

Figura 6.7: Duración de audios

Finalmente, el material de pronunciación tiene un tiempo máximo de tres minutos y dos segundos, mientras que, su tiempo mínimo es de dos segundos. En este tiempo se considera tanto la duración del audio, como el tiempo que le puede tomar al usuario repetirlo, asumiendo que graba su pronunciación por lo menos una vez. La gráfica de los tiempos de las actividades de pronunciación se encuentra representada en la figura 6.8.

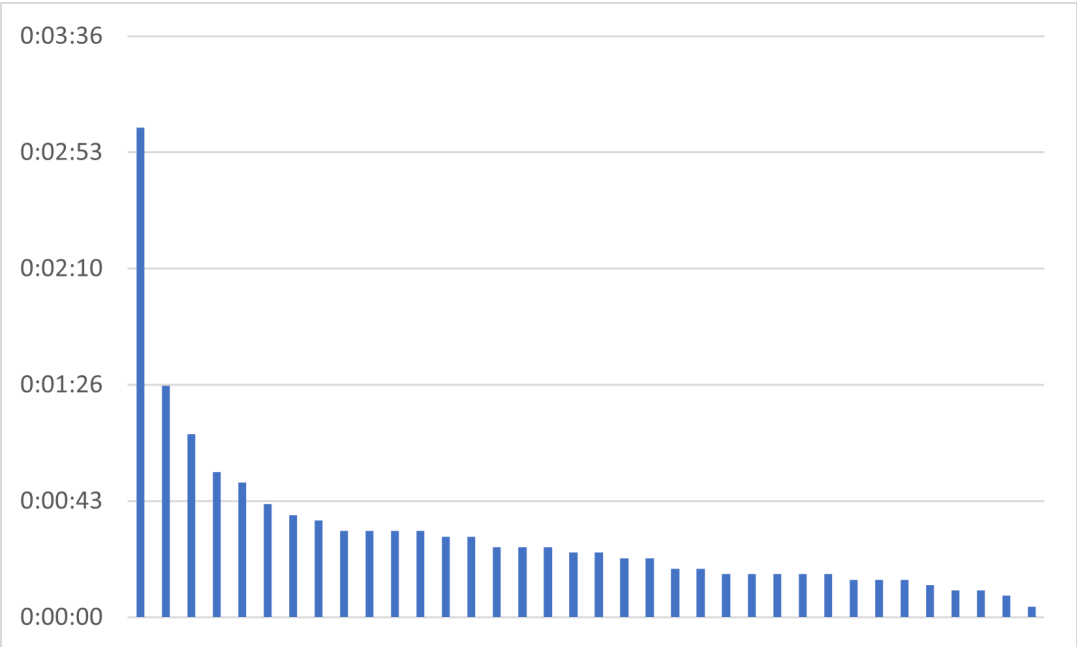

Figura 6.8: Duración de ejercicios de pronunciación

Basándose en la duración promedio del material educativo audiovisual expuesta en las anteriores figuras, se puede determinar que el material de mayor duración es de cinco minutos con cinco segundos. Ahora bien, a diferencia de otros cursos, éste en particular no tiene solamente un tipo de material audiovisual sino tres tipos diferentes. Por ello, es necesario considerarlos grupalmente. Entonces, se consideró una agrupación por sub capítulos, donde la distribución de tiempo es la expuesta en la figura 6.9

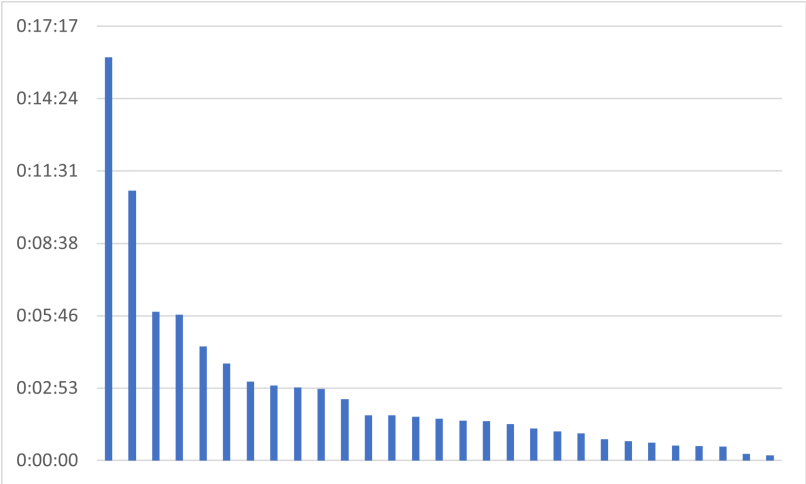

Figura 6.9: Duración de material audiovisual por subcapítulo

A base de la distribución de tiempos por subcapítulos, en donde el subcapítulo con mayor material audiovisual tiene 16 minutos con tres segundos de contenido, y considerando adicionalmente un tiempo de reflexión como el considerado por Barba et al. [38], se opta por un umbral de inactividad de **25 minutos** para determinar el inicio de una nueva sesión en este estudio.

Procrastinación

La procrastinación es un comportamiento en el cual los estudiantes tienden a evitar la realización de tareas hasta el último minuto. Es decir, suspenden el tiempo de aprendizaje hasta que no les quede opción. Este comportamiento se encuentra relacionado con el bajo rendimiento académico en una clase tradicional [39].

Con el fin de relacionar el concepto de sesión de estudios al tiempo de inactividad que se considerara como procrastinación, en este estudio se han analizado dos niveles de comportamiento para este concepto:

1. **Procrastinación intra sesiones.** La cual considera intervalos que contemplen el material audiovisual de mayor duración (cinco minutos y cinco segundos) con su respectivo tiempo de reflexión hasta los 24 minutos y 59 segundos, puesto que, a partir de los 25 minutos se considera ya una nueva sesión.

2. **Procrastinación entre sesiones.** La cual se analiza basándose en el tiempo que el usuario finaliza una sesión de estudio y comienza otra, para ello, se consideran intervalos de tiempo superiores a dos días, puesto que, los fines de semana o días equivalentes a estos no serán considerados dentro del flujo de continuidad del curso. Debido a la naturaleza de los datos de este tipo de procrastinación, su análisis se realizó en una sección independiente posterior a la de la metodología PM^2.

6.2.3. Extracción de datos

La etapa de extracción de datos se dio de forma externa al desarrollo de este estudio. Esta tarea se llevó a cabo en colaboración con la plataforma MOOC basada en Open EDX de la Universidad Politécnica de Valencia administrada por el área de sistemas de la información y las comunicaciones. La tarea se llevó a cabo gracias a la gestión del Doctor Cèsar Ferri quien realizó la solicitud de los datos, obteniendo así cinco archivos en formato CSV, descritos a continuación:

1. Archivo que contiene la información básica del curso que concierne a este caso de estudio.

2. Archivo que contiene la información demográfica de los usuarios del curso.

3. Archivo que contiene la información de aprobación y/o finalización del curso de los usuarios

4. Archivo que contiene la información de inscripción del usuario en el curso.

5. Archivo que contiene los registros de navegación de los usuarios entre los componentes del curso.

6.2.4. Procesamiento de datos

Los datos de los archivos CSV descritos anteriormente fueron procesados mediante el lenguaje de programación R en su IDE RStudio en su versión de escritorio. La limpieza de los datos tuvo dos pasos principales: limpieza de datos de usuarios y limpieza del registro de eventos.

Limpieza de datos de usuarios.

Para la limpieza de datos de usuarios se procedió en tres partes: Eliminación de usuarios no relevantes en el estudio (sin información demográfica y sin actividad en el curso), limpieza de columnas innecesarias y unificación de registros demográficos, registros de inscripción y registros de finalización.

```
#Eliminación de usuarios que no cuentan con la información mínima requerida
descartados <- setdiff(intersect(filter(registro, registro$user_id_encrypt
    %in% setdiff(registro$user_id_encrypt,usuarios$user_id_encrypt))$user_
    id_encrypt,filter(registro, registro$user_id_encrypt%in%setdiff(
    registro$user_id_encrypt,resultados$user_id_encrypt))$user_id_encrypt),
    unique(datos_iniciales$user_id_encrypt))
registro <- filter(registro, !(registro$user_id_encrypt %in% descartados))
#Eliminación de columnas innecesarias de los datos de usuarios
usuarios <- subset(usuarios, select = -c(language))
registro <- subset(registro, select = -c(platform_id,course_id,
    administrator,created))
resultados <- subset(resultados, select = -c(created,modified,course_id,
    platform_id,course_edited_timestamp,course_version,grading_policy_hash,
    passed_timestamp))
#Unión de los archivos de usuarios
datos <- merge(usuarios,registro,by="user_id_encrypt",all = T)
datos <- merge(datos,resultados,by="user_id_encrypt",all = T)
```

Algoritmo 6.1: Procesamiento y unificación de datos de usuarios

La información unificada de los usuarios brinda la posibilidad de identificar grupos determinados de usuarios basándose en sus características en el curso. Principalmente, los usuarios se dividen en dos grupos generales: usuarios certificados y usuarios no certificados. Los usuarios certificados se caracterizan por haber pagado por poder acceder al certificado del curso. En su registro, el campo "mod" tiene el valor "verified". Los usuarios no certificados, también llamados usuarios auditores, se caracterizan por haber accedido al contenido gratuito del curso y no haber pagado por el acceso al certificado. En su registro el campo "mod" tiene el valor "audit".

```
usuarios_certificados <- filter(datos,datos$mod == 'verified')
usuarios_auditores <- filter(datos,datos$mod == 'audit')
```

Algoritmo 6.2: Identificación de grupos de usuarios

Los usuarios certificados requieren una calificación mínima para obtener el certificado de aprobación del curso. En este caso, este valor es de 8 o superior, por lo que los usuarios certificados pueden dividirse en tres grupos: aprobados, cuya nota es de 8 o superior; reprobados, cuya nota es menor a 8 y retirados, que no tienen valor alguno de calificación.

Por otra parte, los usuarios no certificados registran un valor no nulo en el campo de calificación si terminaron el curso, pero si lo abandonaron este valor es nulo.

```
usr_cert_apr <- filter(usuarios_certificados,usuarios_certificados$percent_
    grade>0.79)
usr_cert_rep <- filter(usuarios_certificados,usuarios_certificados$percent_
    grade<0.8,!is.na(usuarios_certificados$percent_grade))
usr_cert_ret <- filter(usuarios_certificados,is.na(usuarios_certificados$
    percent_grade))
usr_audt_fin <- filter(usuarios_auditores,!is.na(usuarios_auditores$percent
    _grade))
usr_audt_ret <- filter(usuarios_auditores,is.na(usuarios_auditores$percent_
    grade))
```

Algoritmo 6.3: Identificación de grupos específicos de usuarios

Limpieza de registro de eventos

Para la limpieza del registro de eventos se procedió en tres partes: identificación de la estructura del curso basándose en la tabla 5.1., eliminación de registros de evento no relacionados y eliminación de información innecesaria.

La columna "module_id" es una columna compuesta conformada por tres partes separadas por el símbolo @. En ésta, primero nos encontramos con el identificador del curso, seguido por el tipo de módulo al que pertenece el evento y, finalmente, el identificador del módulo. En este escenario, sólo se requiere de este último, por lo que se procede también con la eliminación de los dos primeros segmentos de la columna.

```
#Lectura de archivo de estructura del curso
estructura_datos <- read.csv("Estructura del curso.csv",sep=";")
names(estructura_datos)[1]<-"module_id"
#Eliminación de eventos basura y columnas innecesarias
datos_iniciales$module_id <- str_split_fixed(datos_iniciales$module_id, "@",3)
    [,3]
unused_ids <- list("fc53274c8cef4713bdbb12efa29c1401","5
    a70d8a05fbc4473b2a325022f894697","5cade704a181489985a78591426f6bc0","
    e1149d4609ee44a4b7466f0ac9fac39a","e412fcc5114f474e922d18d4c58bc56a")
datos_iniciales <- datos_iniciales[!datos_iniciales$module_id %in% unused_ids,]
#Unión del registro de eventos con la estructura del curso
datos_iniciales <- merge(datos_iniciales,estructura_datos, by.x = "module_id",
    by.y = "module_id", all.x = TRUE)
```

```
#Eliminación de columnas innecesarias del registro de eventos
datos_iniciales <- subset(datos_iniciales,select=-c(module_type.y,course_id,
    platform_id,grade,max_grade,done,created,attempts,watched_time))
```

Algoritmo 6.4: Limpieza del registro de eventos

En base al registro de eventos limpio y los grupos de usuarios identificados anteriormente, se procede a generar los registros de eventos iniciales por cada grupo de usuarios. A partir de estos se pueden generar los modelos de proceso para responder a las preguntas de investigación.

```
eventos_usr_cert_apr <- filter(datos_iniciales,datos_iniciales$user_id_encrypt
    %in% usr_cert_apr$user_id_encrypt)
eventos_usr_cert_rep <- filter(datos_iniciales,datos_iniciales$user_id_encrypt
    %in% usr_cert_rep$user_id_encrypt)
eventos_usr_cert_ret <- filter(datos_iniciales,datos_iniciales$user_id_encrypt
    %in% usr_cert_ret$user_id_encrypt)
eventos_usr_audt_fin <- filter(datos_iniciales,datos_iniciales$user_id_encrypt
    %in% usr_audt_fin$user_id_encrypt)
eventos_usr_audt_ret <- filter(datos_iniciales,datos_iniciales$user_id_encrypt
    %in% usr_audt_ret$user_id_encrypt)
```

Algoritmo 6.5: Registros de eventos por grupos de usuarios

6.2.5. Registro de eventos

El registro de eventos requiere un procesamiento en el cual se identifican las sesiones de usuario y el comportamiento que tuvieron estos en el curso. Para ello, se definió tres funciones: función filtro por nivel, función de identificación de sesiones y función de comportamiento.

1. **Función filtró por nivel.** La función de filtro por nivel extrae los eventos de un solo nivel específico de la estructura jerárquica del curso definida en la figura 6.3.

```
    funcion_filtro=function(eventos,nivel){return(filter(eventos,eventos$
        jerarquia==nivel))}
```

Algoritmo 6.6: Función filtro por nivel

2. **Función de identificación de sesiones.** La función de identificación de sesiones enumera el registro de eventos de los usuarios basándose en el umbral de inactividad de 25 minutos definido anteriormente. Por cada pausa de 25 minutos o superior la función enumera los siguientes eventos como una nueva sesión.

```
    funcion_sesiones=function(eventos,umbral){return(eventos%>% group_by
        (user_id_encrypt)%>% arrange(modified,.by_group=TRUE)%>% mutate(
        timeValue=difftime(modified,lag(modified),units="secs"))%>%
        mutate(timeValue=replace_na(timeValue,0))%>% mutate(code=case_
        when(row_number()==1~1,timeValue>(umbral*60)~1))%>% mutate(sesion
        =cumsum(!is.na(code))))}
```

Algoritmo 6.7: Función de identificación de sesiones

3. **Función de comportamiento.** La función de comportamiento identifica los periodos de inactividad dentro de una sesión, que no son lo suficientemente largos como para pasar el umbral de los 25 minutos y ser considerados una nueva sesión. En este caso, se consideran tiempos mayores a los 10 minutos.

```
funcion_comportamiento=function(eventos,umbral){return(eventos%>%
    group_by(user_id_encrypt,sesion)%>% arrange(modified,.by_group=
    TRUE)%>% mutate(nivel_atencion=ifelse(lead(timeValue)>60*umbral,'
    Distraccion',''))%>% mutate(nivel_atencion = replace_na(nivel_
    atencion,'')))}
```

Algoritmo 6.8: Función de comportamiento

6.2.6. Modelos de procesos

En esta etapa se procede con la aplicación del algoritmo de minería de procesos para descubrir el modelo de procesos. Esta acción permite extraer un modelo real de forma automática. Para este fin se utiliza la herramienta de software Disco Fluxicon, la cual aplica un algoritmo adaptado de FuzzyMiner [40], que se caracteriza por adaptarse a escenarios flexibles. El procesamiento de los registros de eventos se realiza con respecto a las necesidades requeridas para responder a las preguntas de investigación.

RQ1. ¿Existen variaciones en el comportamiento de los diferentes grupos de participantes en un curso MOOC?

Para dar respuesta a esta pregunta se considera la navegación a través de los módulos de tipo "Chapter" cuyo nivel de jerarquía es de 2 con el fin de determinar el desplazamiento general a través del material del curso. Para ello, se utilizó la función del algoritmo 6.6 generando así cinco registros de eventos, uno por cada grupo de usuarios.

module_type	modified	user_id_encrypt	jerarquia	Nombre
chapter	21/08/2020 0:02	4353415650554450	2	Mid-term exam
chapter	18/05/2020 23:26	435743525c504853	2	Mid-term exam
chapter	20/04/2020 15:12	42524a5158584554	2	Mid-term exam
chapter	24/05/2020 6:31	4253475b5f574952	2	Mid-term exam
chapter	06/08/2020 13:03	4350435a5e554156	2	Mid-term exam
chapter	10/04/2020 14:26	425c44565e57415c	2	Mid-term exam
chapter	25/05/2020 5:16	4354475b5e594550	2	Mid-term exam

Tabla 6.13: Muestra del registro de eventos

Para generar los modelos de procesos es necesario cargar los registros de eventos en el software Disco. Con el registro de eventos cargado se procedió con la identificación de las variables mínimas requeridas por el software. La identificación de las variables mínimas requeridas consiste en asignar una o varias columnas del registro de eventos a cada una de las variables. Estas variables son descritas en la tabla 6.14 con su relación con el registro de eventos que se analizó. Por otra parte, en la figura 6.10 se puede ver la interfaz de asignación de columnas a las variables mínimas requeridas.

Variable	Columna
Case Id	user_id_encrypt
Activity	Nombre
Timestamp	modified

Tabla 6.14: Identificación de variables – RQ1

Remove

	module_type.x	modified	user_id_encrypt	jerarquia	Nombre
aaa79c230	chapter	2020-08-21 00:02:53	4353415650554450	2	Mid-term exam
aaa79c230	chapter	2020-05-18 23:26:07	435743525c504853	2	Mid-term exam
aaa79c230	chapter	2020-04-20 15:12:22	42524a5158584554	2	Mid-term exam
aaa79c230	chapter	2020-05-24 06:31:59	4253475b5f574952	2	Mid-term exam
aaa79c230	chapter	2020-08-06 13:03:52	4350435a5e554156	2	Mid-term exam
aaa79c230	chapter	2020-04-10 14:26:53	425c44565e57415c	2	Mid-term exam
aaa79c230	chapter	2020-05-25 05:16:37	4354475b5e594550	2	Mid-term exam
aaa79c230	chapter	2020-07-25 22:22:09	435045505c564250	2	Mid-term exam
aaa79c230	chapter	2020-06-24 18:47:47	4357455b5b504055	2	Mid-term exam
aaa79c230	chapter	2020-05-16 14:51:44	425c4a5b5a574556	2	Mid-term exam
aaa79c230	chapter	2020-04-08 14:28:44	425d46565b524953	2	Mid-term exam
aaa79c230	chapter	2020-04-26 17:22:41	4357435358514553	2	Mid-term exam

Figura 6.10: Identificación de variables requeridas por Disco

Los modelos de procesos obtenidos se muestran en las figuras 6.11 para los usuarios no certificados y en la figura 6.12 para los usuarios certificados.

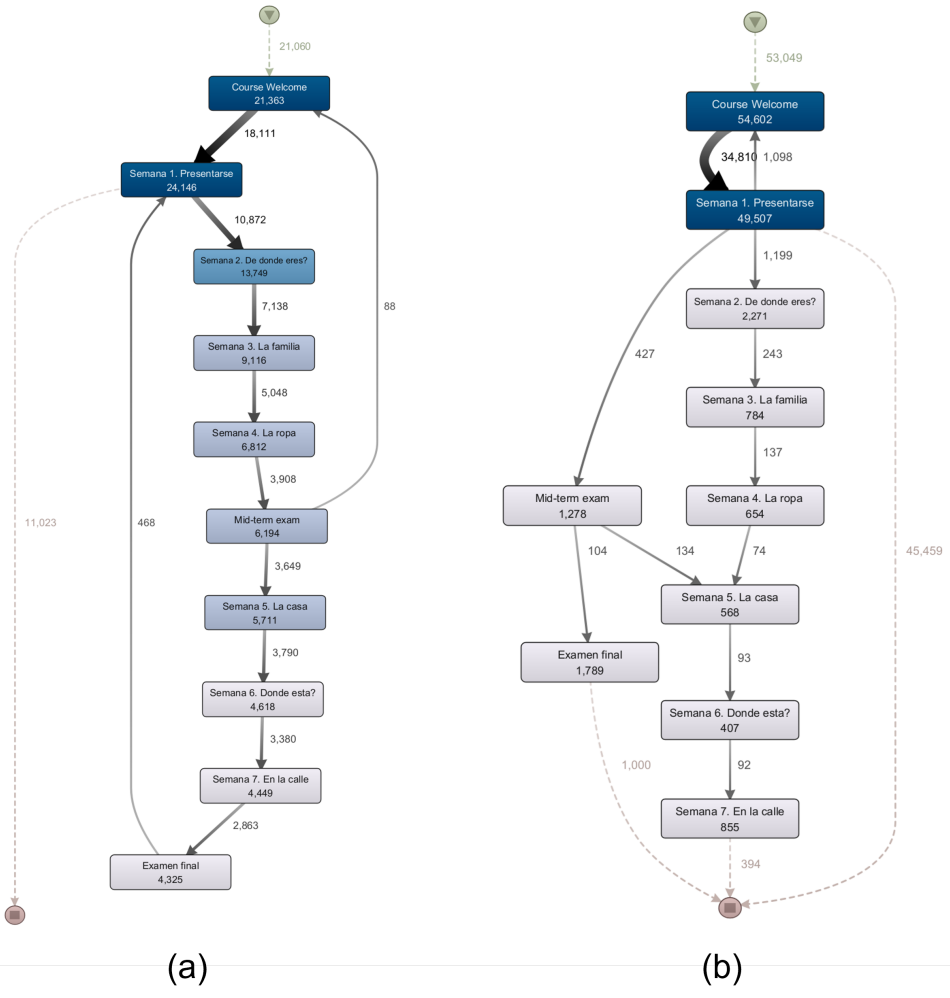

Figura 6.11: Modelo de procesos para los usuarios no certificados (a) que finalizaron (b) que se retiraron.

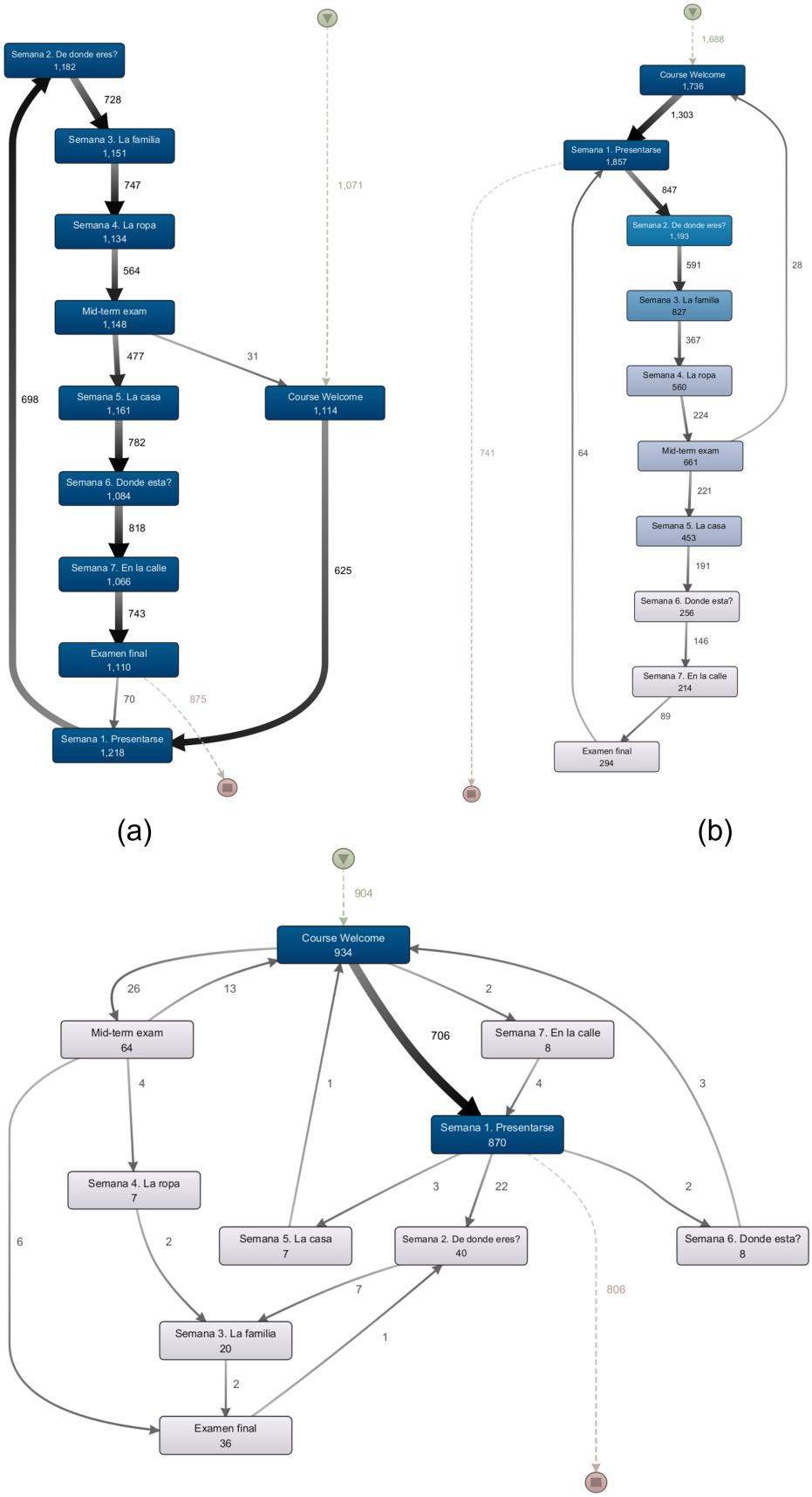

Figura 6.12: Modelo de procesos para los estudiantes certificados (a) aprobados (b) reprobados (c) retirados

RQ2. ¿Cuáles son las principales características del comportamiento de los diferentes grupos de participantes en un curso MOOC?

Para responder a esta pregunta de investigación se procedió con la identificación del registro de eventos con sesiones a partir de dos niveles jerárquicos: nivel 4 y nivel 5. El nivel 4 corresponde a los módulos de tipo "vertical" los que actúan como un contenedor de las diferentes lecciones existentes en el curso, mientras que el nivel 5 corresponde a los módulos "video", "problem" y "drag-and-drop".

Para los dos niveles jerárquicos se generaron los registros de eventos haciendo uso de las funciones expuestas en los algoritmos 6.6 y 6.7.

La tabla 6.15 contiene una muestra del registro de eventos a nivel jerárquico 4. La asignación de las columnas del registro de eventos a las variables mínimas requeridas por la herramienta de software se muestran en la tabla 6.16. En este caso, la variable "Case Id" se compone de dos columnas: user_id_encrypt y sesión.

modified	user_id_encrypt	jerarquia	Tipo_contenido	sesion
26/09/2020 1:23	415141575d564253	4	Pronunciacion	1
04/09/2020 17:04	415142545f5448	4	Video	1
01/04/2020 12:51	4151465a5955405d	4	Video	1
22/04/2020 5:43	4151465a5955405d	4	Video	2
03/05/2020 5:54	4151465a5955405d	4	Video	3
03/05/2020 6:32	4151465a5955405d	4	Video	4

Tabla 6.15: Muestra del registro de eventos por sesiones a nivel jerárquico 4.

Variable	Columna
Case Id	user_id_encrypt & sesion
Activity	tipo_contenido
Timestamp	modified

Tabla 6.16: Identificación de variables – RQ2 - Nivel 4

Los modelos de procesos resultantes se encuentran en las figuras 6.13 para los usuarios no certificados y en la figura 6.14 para los usuarios certificados.

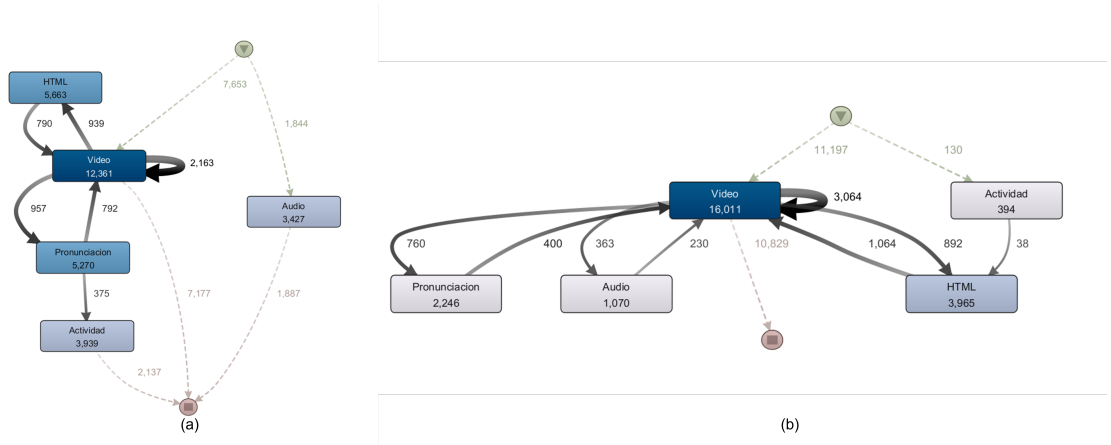

Figura 6.13: Modelo de procesos de para los usuarios no certificados (a) que finalizaron (b) que abandonaron

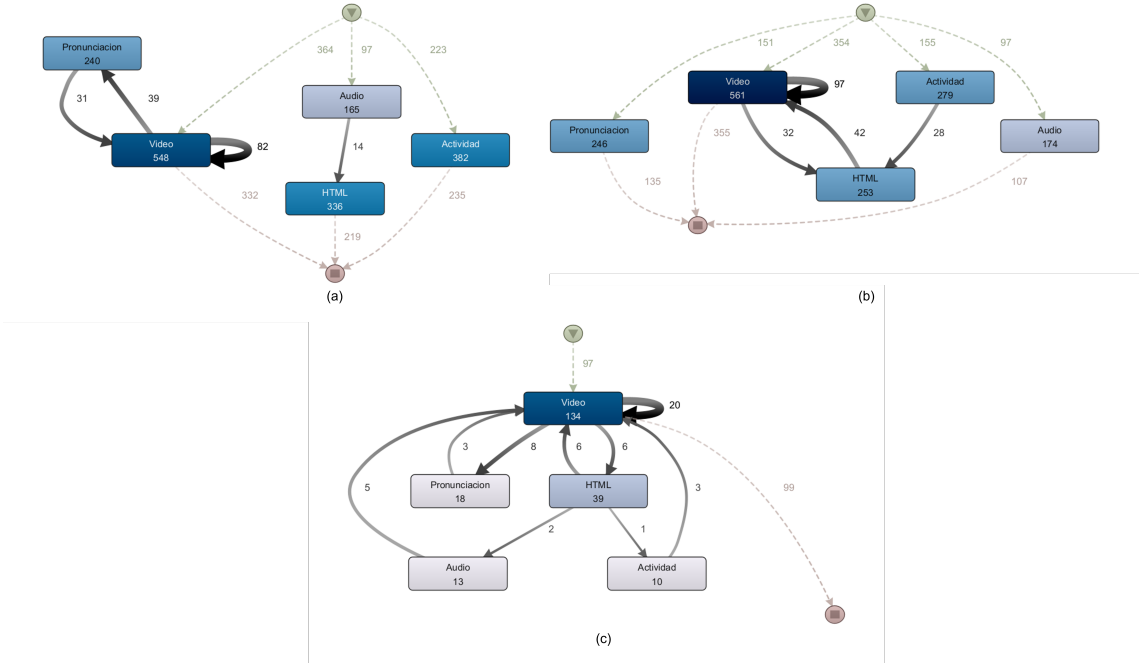

Figura 6.14: Modelo de procesos de para los usuarios certificados (a) aprobados (b) reprobados (c) retirados

Una muestra del registro de eventos a nivel jerárquico 5 se detalla en la tabla 6.17. La distribución de las columnas en las variables mínimas requeridas requiere que la variable "Case Id" tenga asignada dos columnas: user_id_encrypt y sesión. La asignación de las columnas a las variables mínimas requeridas para este registro de eventos se expone en la tabla 6.18.

module_type	modified	user_id_encrypt	jerarquia	sesion
problem	20/04/2020 6:11	415040575c524657	5	1
problem	20/04/2020 6:11	415040575c524657	5	1
video	20/04/2020 6:13	415040575c524657	5	1
problem	20/04/2020 6:50	415040575c524657	5	2
video	20/04/2020 7:36	415040575c524657	5	3
drag-and-drop-v2	20/04/2020 8:13	415040575c524657	5	4
drag-and-drop-v2	20/04/2020 8:15	415040575c524657	5	4

Tabla 6.17: Muestra del registro de eventos por sesiones a nivel jerárquico 5

Variable	Columna
Case Id	user_id_encrypt & sesion
Activity	module_type
Timestamp	modified

Tabla 6.18: Identificación de variables – RQ2 - Nivel 5

A partir de estos registros de eventos, se procede con la generación de los modelos de proceso por sesiones para los estudiantes certificados que aprobaron y reprobaron, además de los estudiantes no certificados que finalizaron el material. Los estudiantes certificados y no certificados que abandonaron el curso solo tienen interacciones con videos

a este nivel jerárquico, por lo tanto, sus modelos de procesos son omitidos. Los modelos de procesos obtenidos se encuentran en la figura 6.15.

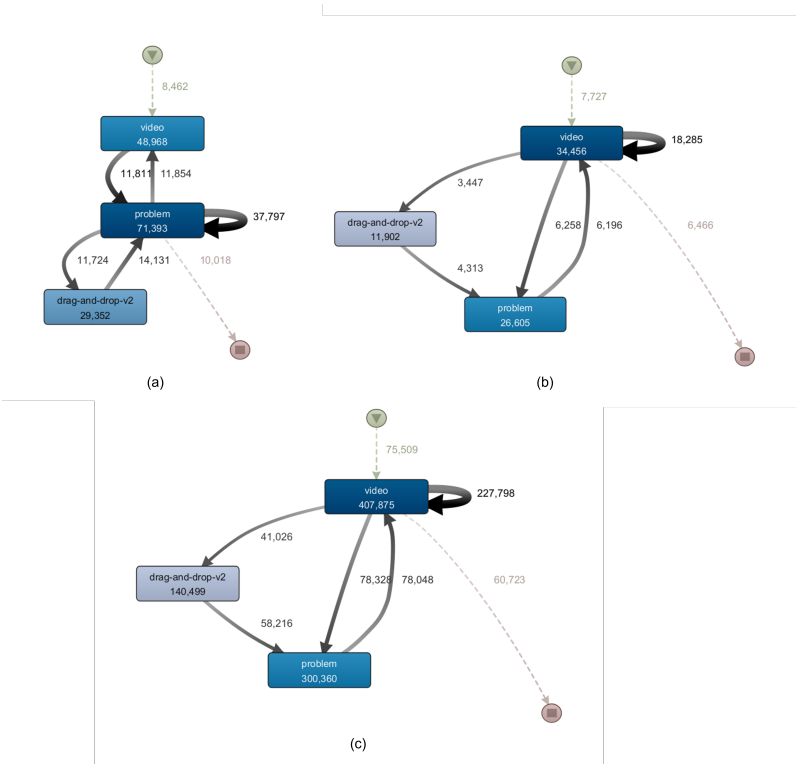

Figura 6.15: Modelo de procesos para los usuarios certificados (a) aprobados (b) reprobados y usuarios no certificados (c) que finalizan la revisión del material

RQ3. ¿Existe relación entre la procrastinación y el éxito o fracaso de un participante en un curso MOOC?

Esta pregunta de investigación requiere identificar los instantes en que el usuario hace pausas de 10 minutos o más en una sesión de estudio. Para ello se tomaron eventos de nivel jerárquico 4. Se utilizaron las funciones de los algoritmos 6.6, 6.7 y 6.8 para seleccionar los eventos, definir las sesiones de estudio e identificar los momentos de distracción respectivamente. En la tabla 6.19 se presenta una muestra del registro de eventos.

modified	user_id_encrypt	jerarquia	Tipo_contenido	sesion	atencion
01/04/2020 8:19	415240525f56495c	4	Pronunciacion	1	
01/04/2020 9:50	415240525f56495c	4	Video	2	
01/04/2020 12:25	415240525f56495c	4	Video	3	Distr
01/04/2020 12:45	415240525f56495c	4	Audio	3	
01/04/2020 21:00	415240525f56495c	4	Pronunciacion	4	Distr

Tabla 6.19: Muestra del registro de eventos por sesiones

Las variables mínimas requeridas en este escenario están conformadas por dos variables compuestas y una simple. La distribución se detalla en la tabla 6.20.

Variable	Columna
Case Id	user_id_encrypt & sesion
Activity	tipo_contenido & atencion
Timestamp	modified

Tabla 6.20: Identificación de variables – RQ2 - Nivel 5

A partir de estos registros de eventos se generó los modelos de proceso de procrastinación intra-sesiones reflejados en las figuras 6.16 y 6.17.

La composición de la variable "activity" permite identificar los puntos de procrastinación en los modelos de procesos. Las actividades que tienen la palabra "distracción" representan un punto en que el usuario finalizó esa actividad, pero se tardó 10 minutos o más en pasar a la siguiente.

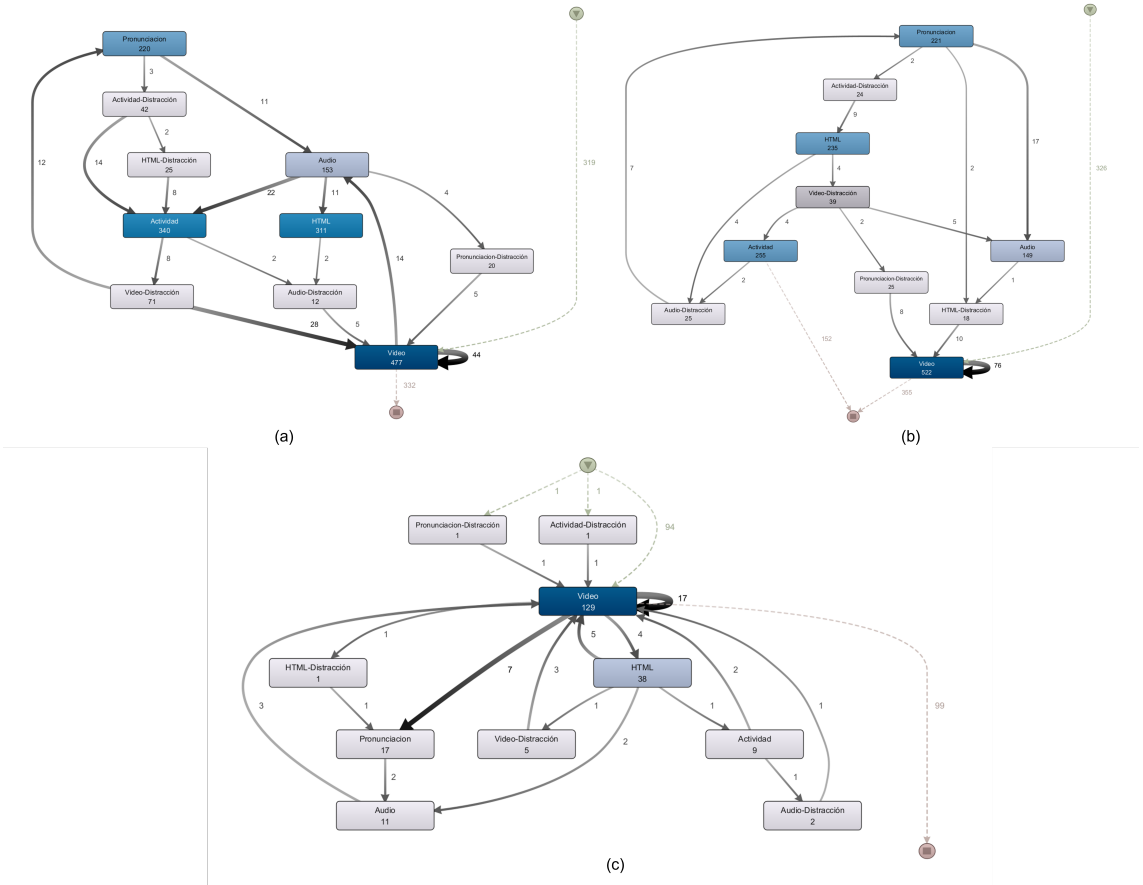

Figura 6.16: Modelo de procesos de para los usuarios certificados (a) aprobados (b) reprobados (c) retirados

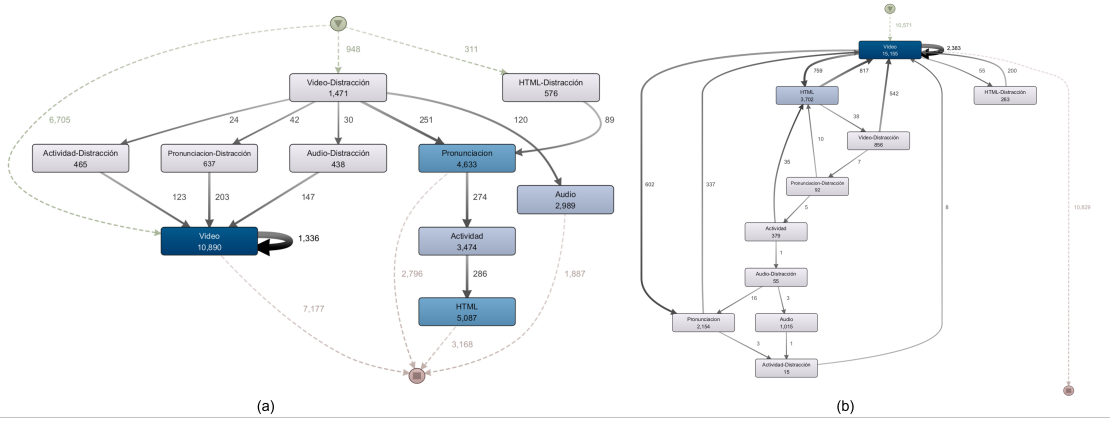

Figura 6.17: Modelo de procesos de para los usuarios no certificados (a) que finalizaron (b) que abandonaron

CAPÍTULO 7
Resultados

Para este capítulo se realizó la última etapa de la metodología PM2, la etapa de evaluación de modelos. En esta etapa se interpretaron los modelos de procesos obtenidos. A partir de la cual, se obtienen los siguientes resultados. Para cada resultado se especifica a qué pregunta de investigación pertenece.

R1. El abandono se produce mayoritariamente a partir de la segunda semana (RQ1)

En contraste con la figura 6.2 del flujo de usuarios, en la figura 7.1 se puede observar que los usuarios abandonan el curso a partir de la segunda semana. Ya sean usuarios certificados o no certificados, tienden a revisar el contenido de la semana de bienvenida y de la semana de presentación.

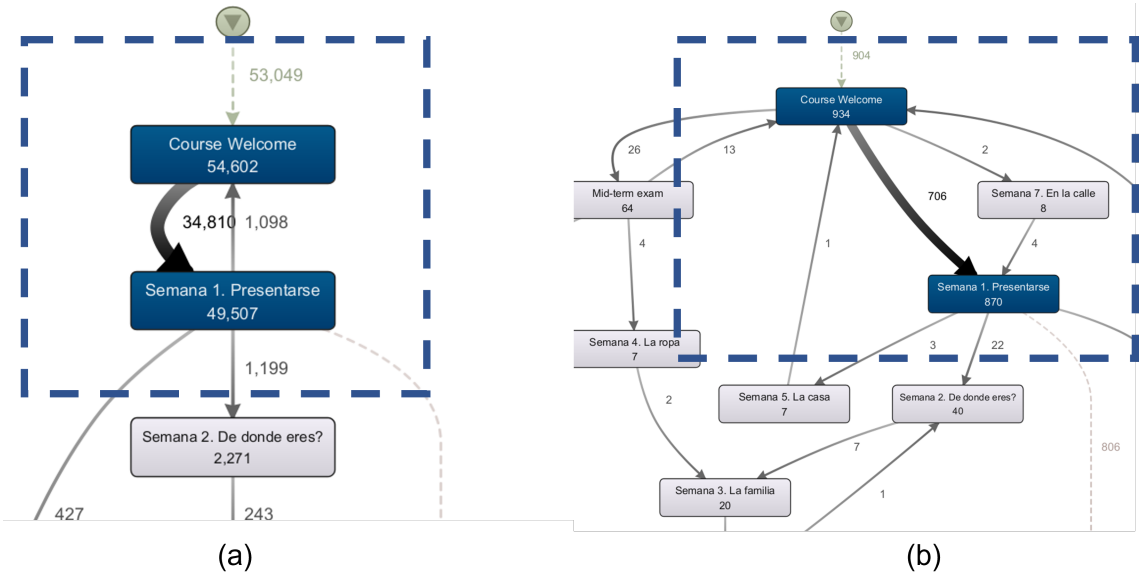

Figura 7.1: Interacción de usuarios (a) no certificados (b) certificados retirados con el material del curso en capítulos.

Por otra parte, como se puede observar en la figura 7.2, los estudiantes no certificados que abandonaron el curso acceden a las dos primeras semanas en un tiempo promedio de 16 minutos, mientras que, los estudiantes certificados lo hacen en un tiempo promedio de aproximadamente dos horas. Esto evidencia que los usuarios certificados que abandonaron el curso desde un inicio no tenían interés en el contenido, mientras que, los usuarios no certificados perdieron la motivación luego de la segunda semana.

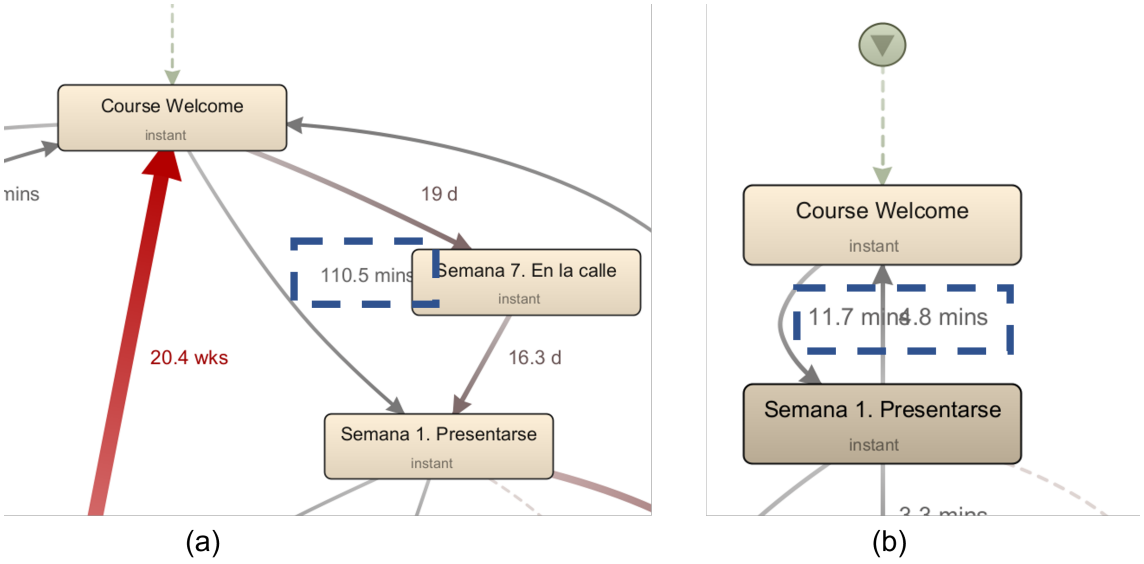

Figura 7.2: Tiempo promedio de desplazamiento entre los primeros capítulos para usuarios (a) certificados (b) no certificados que abandonaron el curso.

R2. Los usuarios certificados que abandonaron el curso buscaron contenido específico en el material (RQ1)

En la figura 7.3 se puede observar un comportamiento no secuencial de los usuarios certificados que abandonaron el curso. En este comportamiento, la bienvenida y la semana de introducción son los únicos capítulos que fueron revisados en el orden establecido. Mientras que el desplazamiento de los usuarios por los otros capítulos es aleatorio.

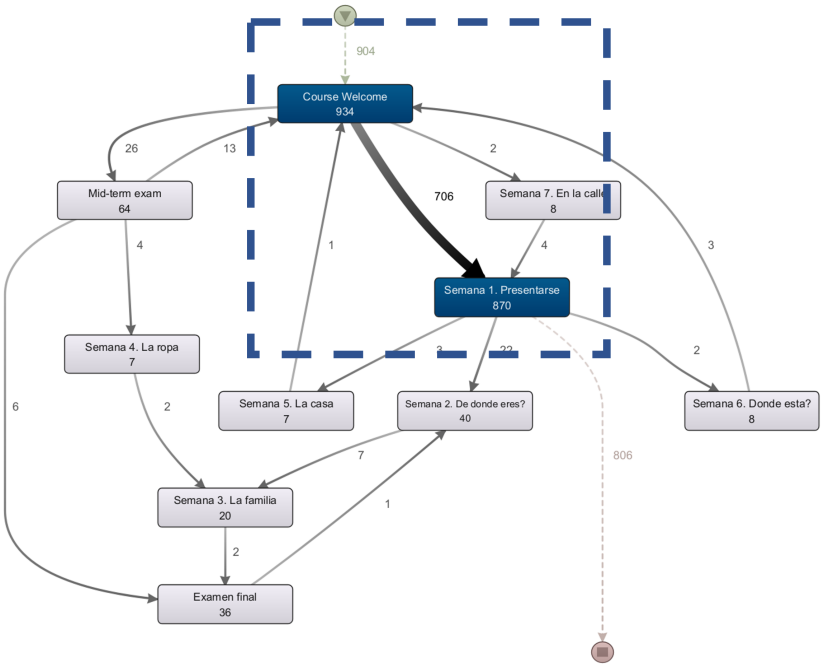

Figura 7.3: Comportamiento de usuarios certificados que abandonaron el curso en el ámbito de capítulos.

Ahora bien, en la figura 7.4 se puede observar que los usuarios certificados que abandonaron el curso tienen tiempos de inactividad entre capítulos que superan enormemente la duración del curso. En donde, el mayor tiempo promedio es de 20 semanas. Mientras tanto, el primer examen calificado no pasa de los seis minutos de acceso hasta cambiar de capítulo. Basándose en este comportamiento se deduce que estos usuarios pagaron por el certificado para poder acceder al material de curso siempre que lo necesiten, sin tener que preocuparse porque la plataforma limite su acceso.

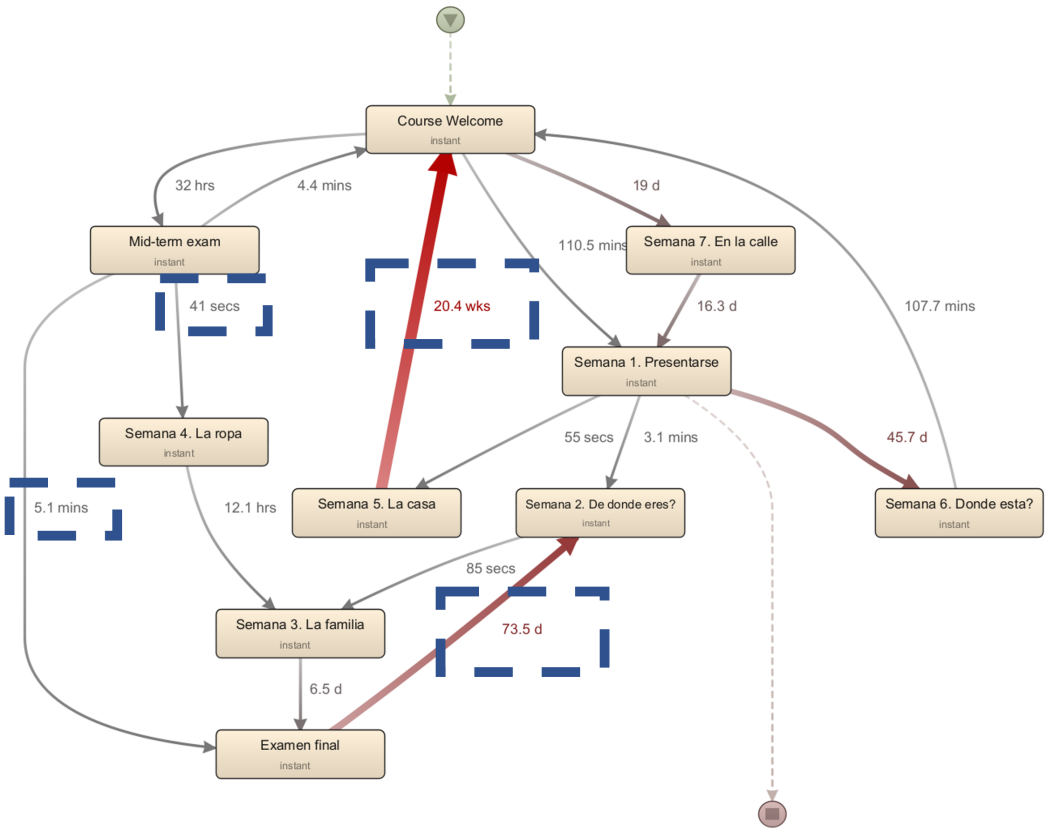

Figura 7.4: Tiempos de desplazamiento entre capítulos para usuarios certificados que abandonaron el curso.

R3. El abandono de los usuarios no certificados se da al contrastar que no pueden realizar las evaluaciones calificadas (RQ1)

Los usuarios no certificados que abandonan el curso tienen un comportamiento que se aproxima a la estructura de este. En la figura 7.5 (a) se puede observar que existe una ramificación en la cual, un grupo de usuarios se desplazan desde la evaluación intermedia directamente a la evaluación final antes de finalizar su recorrido por el contenido. Esto, en contraste con la figura 7.5 (b), en donde se observa que este desplazamiento se realiza en menos de un minuto evidencia una verificación de falta de acceso al contenido de evaluación calificado. A partir de esta observación, se puede decir que una de las razones de abandono del curso por parte de estudiantes no certificados es la falta de acceso al contenido de evaluación calificado, puesto que, varios casos finalizan su recorrido luego de esta verificación.

Por otra parte, la figura 7.5 (b) evidencia tiempos relativamente cortos entre capítulos a partir de la segunda semana. A partir de lo que se puede inferir que estos usuarios no certificados revisaron el curso brevemente antes de abandonarlo.

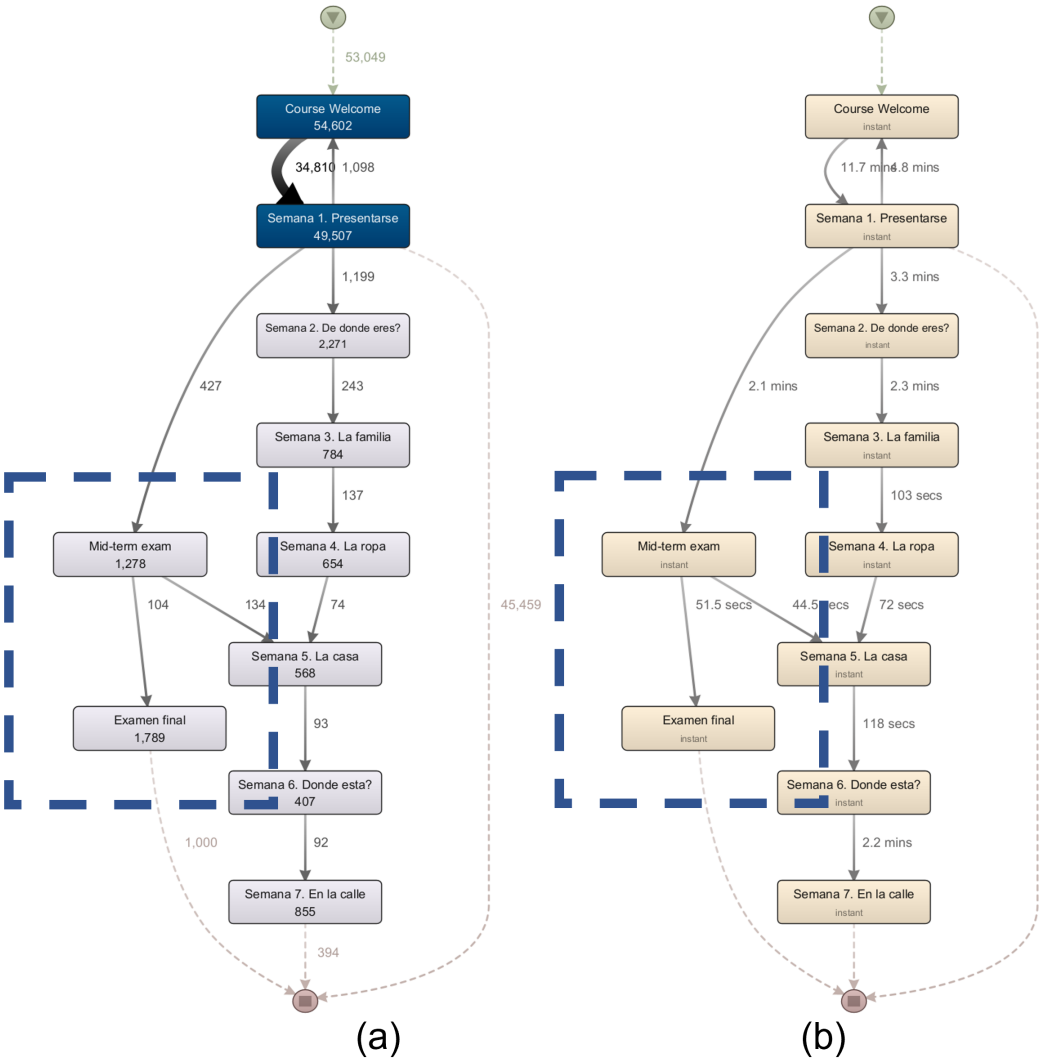

Figura 7.5: Comportamiento de usuarios no certificados que abandonaron el curso con base en (a) transiciones (b) tiempos de desplazamiento.

R4. Los usuarios certificados tienden a regresar en el contenido habiendo finalizado las evaluaciones calificadas (RQ1)

Los usuarios certificados que no abandonaron el curso tienen la tendencia de regresar al contenido en el momento de llegar a una evaluación calificada. Como se puede observar en la figura 7.6, tanto usuarios certificados aprobados como usuarios certificados reprobados regresan en el material al llegar a una evaluación, pero, los aprobados realizan esta acción solamente con la evaluación intermedia. Este comportamiento evidencia un proceso de recapitulación del contenido posterior a una evaluación, en donde, los usuarios aprobados al ver que obtuvieron el certificado no sienten la necesidad de recapitular el contenido. Por otra parte, los usuarios reprobados presentan la necesidad de recapitular el contenido al ver que no obtuvieron la calificación mínima requerida.

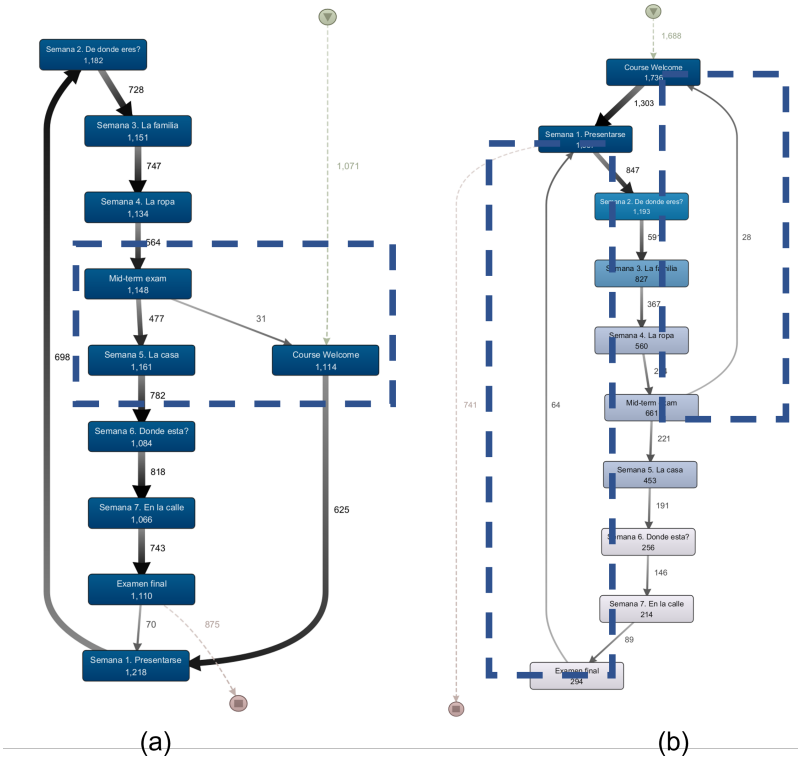

Figura 7.6: Comportamiento de usuarios certificados (a) aprobados (b) reprobados.

R5. Las sesiones de estudio de los usuarios certificados aprobados contemplan un recorrido complementario entre el material y las evaluaciones. (RQ2)

Los usuarios certificados que aprobaron el curso tienden a realizar sesiones que contemplan por lo menos dos componentes en el nivel jerárquico 5. Como se puede observar en la figura 7.7 (a), los usuarios aprobados comienzan sus sesiones de estudio con videos, pero, involucran los dos tipos de actividad existente en el curso. Mientras que, los usuarios reprobados realizan sesiones de estudio que mayoritariamente involucra solamente videos.

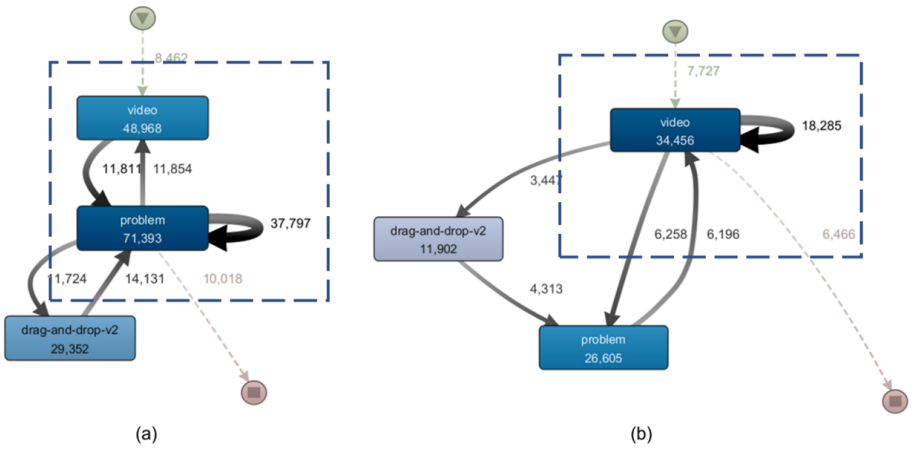

Figura 7.7: Comportamiento de usuarios certificados (a) aprobados (b) reprobados con respecto a sesiones de estudio.

R6. Los usuarios certificados que reprobaron tienen a buscar las respuestas de las evaluaciones en el contenido antes de responder (RQ2).

Los usuarios certificados aprobados realizan bucles de interacción entre los dos tipos de evaluaciones existentes. En la figura 7.8 (a) se evidencia un comportamiento constante en las evaluaciones de los usuarios aprobados. Ahora bien, los estudiantes reprobados no presentan este comportamiento, como bien queda evidenciado en la figura 7.8 (b), los videos se alternan entre los dos tipos de evaluación durante una sesión. Esto evidencia que los usuarios aprobados intentan no necesitar acceder al material de apoyo para responder a las evaluaciones. Por otra parte, los usuarios reprobados tienden a realizar búsquedas de apoyo en el material para responder a las evaluaciones.

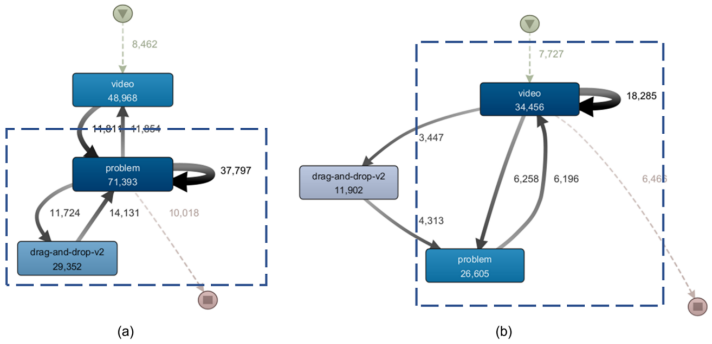

Figura 7.8: Comportamiento de usuarios certificados (a) aprobados (b) reprobados en cuanto a sesiones de estudio.

R7. Los usuarios certificados aprobados realizan sesiones de estudio enfocadas en resolver las evaluaciones (RQ2)

A un nivel jerárquico 4, los usuarios certificados que aprobaron el curso presentan un comportamiento en el que realizan sesiones de estudio enfocadas en resolver las actividades. Tomando en cuenta que los módulos de jerarquía 4 de tipo actividad pueden contener tanto módulos de tipo "problem" como "drag-and-drop" o una combinación de ambos. En contraste con el resultado 6 (R6), en donde se observó secuencias continuas de módulos de actividad. En la figura 7.9 (a) se puede observar este comportamiento mencionado. Por otra parte, los usuarios reprobados involucran otros componentes dentro de una sesión de estudio, cuando esta comenzó con una actividad, como se puede observar en la figura 7.9 (b). A partir de estas observaciones, se puede confirmar lo resaltado en el R6, que los usuarios reprobados intentan encontrar las respuestas de las evaluaciones en el material.

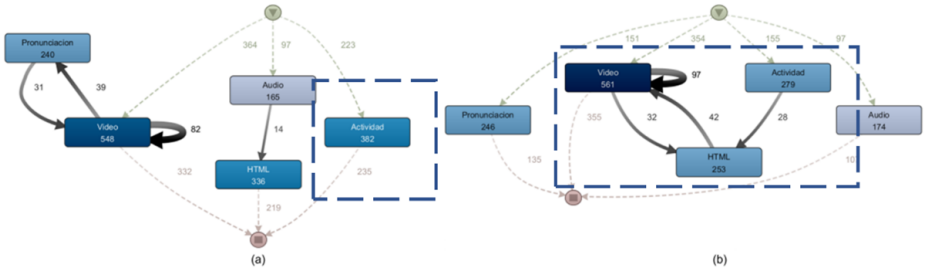

Figura 7.9: Comportamiento de usuarios certificados (a) aprobados (b) reprobados en el nivel jerárquico 4.

R8. Los usuarios certificados aprobados aprovechan de mejor forma el material auditivo y de pronunciación (RQ2)

El material de audio y de pronunciación se encuentra más involucrado en las sesiones de estudio de los estudiantes certificados. Esto se evidencia en la figura 7.10 (a), en donde se puede observar que tanto los audios como los ejercicios de pronunciación forman parte de las sesiones de estudio en combinación con las lecturas y los videos respectivamente. Mientras que, son lecciones aisladas en las sesiones de estudio de los usuarios reprobados.

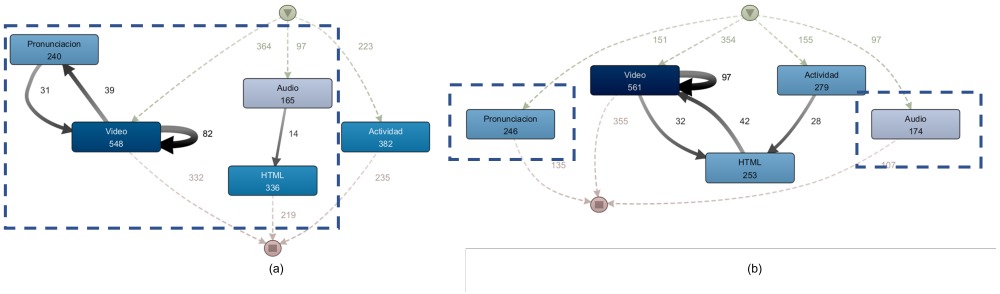

Figura 7.10: Comportamiento de usuarios certificados (a) aprobados (b) reprobados en el nivel jerárquico 4.

R9. Los usuarios no certificados realizan sesiones de revisión de material de lectura y audiovisual antes de acceder a las actividades disponibles, pero tienden a realizar sesiones de audio independientes. (RQ2)

Los usuarios no certificados tienden a realizar sesiones de estudio excluyendo las lecciones de audio. Como se puede observar en la figura 7.11 (a), las sesiones de estudio de los estudiantes no certificados abarcan prácticamente todos los tipos de lecciones disponibles, a excepción de audio, el cual suele ser tratado en sesiones independientes. Como se puede observar en la figura 7.11 (b), las sesiones de audio son sesiones cortas, en las cuales, el usuario accede y se retira sin interactuar con otras lecciones. Basándose en esto, se puede inferir que el audio es el tipo de lección que menos llama la atención de los usuarios no certificados.

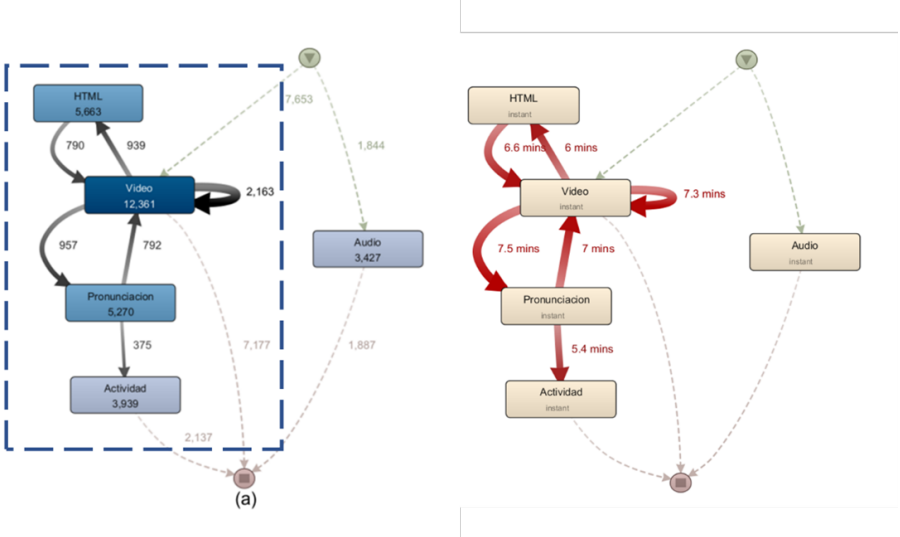

Figura 7.11: Comportamiento de usuarios no certificados que finalizaron el curso.

R10. La procrastinación intra-sesiones no es común para los usuarios que aprobaron (RQ3)

Para los estudiantes que aprobaron existen casos muy puntuales de procrastinación dentro de una sesión de estudio, donde el más común se da entre secuencias de video como se observa en la figura 7.12 (a). Mientras que, para los estudiantes reprobados la procrastinación dentro de una sesión se da más comúnmente cuando el curso requiere cambiar de material para poder continuar. Las pausas son más frecuentes si se requiere ir de un video a una actividad, pero, la falta de interés a las actividades de audio de los estudiantes reprobados se evidencia en la figura 7.12 (b). Donde estos componentes son involucrados también posterior a una distracción al finalizar un video, reduciendo aún más el interés al llegar a actividades de pronunciación.

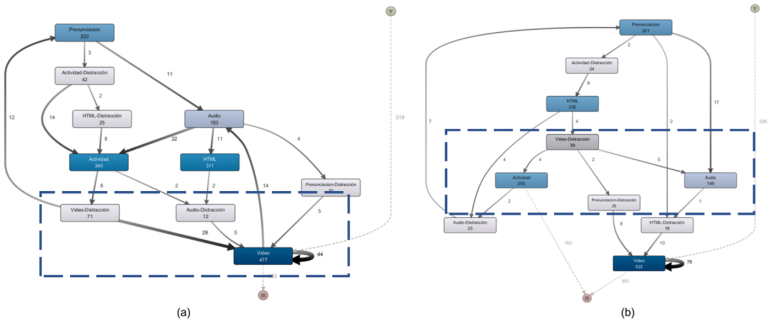

Figura 7.12: Atención de estudiantes certificados aprobados (a) y reprobados (b).

R11. El cambio de tipo de lección tiende a provocar distracción en usuarios no certificados (RQ3)

Sobre la base de los resultados expuestos en la figura 7.13, se puede observar que los videos continuos mantienen la atención de los usuarios. Ahora bien, cuando el flujo del curso requiere desplazarse a otro tipo de lecciones, los usuarios tienden a procrastinar durante la sesión. Un comportamiento peculiar se presenta en las interacciones con los ejercicios de pronunciación, donde, estos ejercicios pierden interés si son anteriores a un video. Pero, despiertan nuevamente el interés cuando anteceden a otras actividades que requieran interacción del usuario.

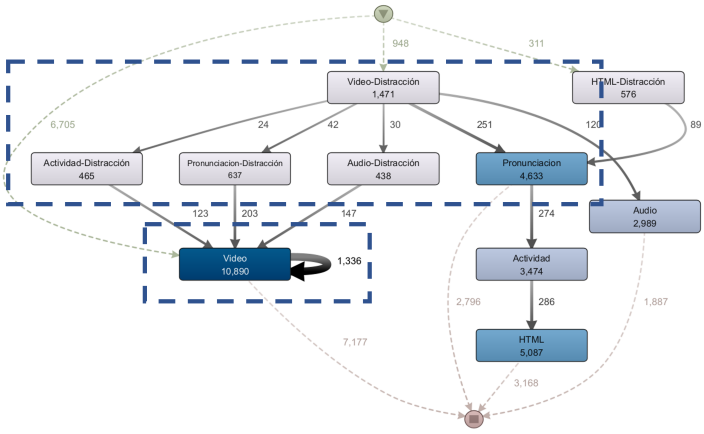

Figura 7.13: Atención de estudiantes no certificados que finalizaron el curso

R12. Las actividades y la pronunciación tienden despertar el interés de los usuarios que procrastinaron durante la sesión de estudio (RQ3).

Las actividades y la pronunciación tienden a comportarse como captadores de interés de los usuarios. Como se puede observar en la figura 7.14, tanto para los estudiantes certificados como no certificados que no abandonaron el curso, al momento de llegar a una actividad que involucra más que ver o escuchar, el nivel de atención aumenta. Ahora, cuando existe distracción luego de una actividad de pronunciación, el tipo de lección que más comúnmente vuelve a captar la atención del usuario son los videos.

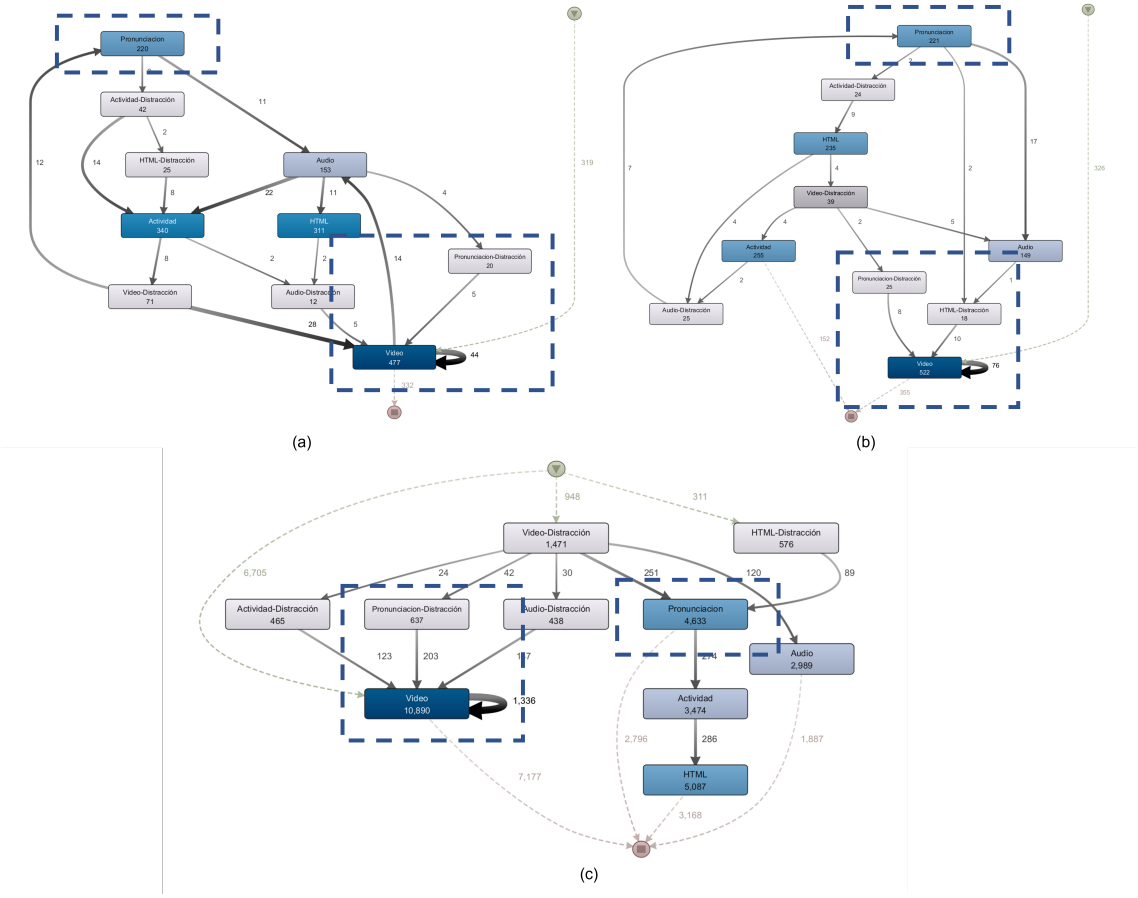

Figura 7.14: Nivel de atención de usuarios certificados (a) aprobados (b) reprobados y usuarios no certificados (c) que finalizaron el curso.

7.1 Análisis de procrastinación entre sesiones

Para la procrastinación entre sesiones se tomó en cuenta los eventos en todos los niveles jerárquicos definidos en la estructura del curso. Con el objetivo de que se consideren todas las interacciones posibles con el curso antes de calcular el tiempo entre las sesiones.

Para el cálculo de tiempo entre sesiones se utilizó la funciones definidas en los algoritmos 6.7 para etiquetar el registro de eventos con las sesiones de los usuarios, y la función 7.1 para calcular el tiempo entre el fin de una sesión y el inicio de otra.

```
funcion_entre_sesiones=function(eventos){return(
subset(eventos %>% group_by(user_id_encrypt,sesion) %>% mutate(inicio=first(
    modified)) %>% mutate(fin=last(modified)),select=c(user_id_encrypt,sesion,
    inicio,fin)) %>% distinct() %>% group_by(user_id_encrypt) %>% mutate(
```

```
    duracion=difftime(fin,inicio,units="mins"))%>% mutate(descanso = difftime(
    lead(inicio),fin,unit="hours"))
)}
```

Algoritmo 7.1: Función de calculo de tiempo entre las sesiones

A partir del cálculo de tiempo entre sesiones se procedió con el análisis de los intervalos de tiempo mediante la generación de funciones de distribución. Para ello los datos fueron segmentados en tres partes:

1. **Descanso esperado:** El cual contempla periodos entre 25 minutos a 48 horas.
2. **Procrastinación ordinaria:** La cual contempla periodos de tiempo que van desde las 48 horas hasta las 336 horas (10 semanas), es decir, el periodo de tiempo que dura el curso.
3. **Procrastinación extraordinaria:** La cual contempla periodos de tiempo desde las 336 horas en adelante.

Esta segmentación se realizó para los tres grupos de usuarios certificados identificados. Ahora bien, con esta segmentación se procedió con la generación de las funciones de distribución obteniendo así los siguientes resultados:

Función de distribución de descanso esperado para usuarios certificados

El comportamiento de los usuarios certificados dentro del Intervalo de tiempo de un descanso esperado es similar tanto para los aprobados como los reprobados como se puede observar en la figura 7.15, en donde, sus respectivas curvas se encuentran más pronunciadas hacia la izquierda. Por otra parte, los usuarios retirados, aunque asemejan el mismo comportamiento, en este se puede encontrar sutiles diferencias, como que, su pico más alto es también, más ancho, evidenciando así intervalos de tiempo más variados en terminar una sesión y comenzar otra.

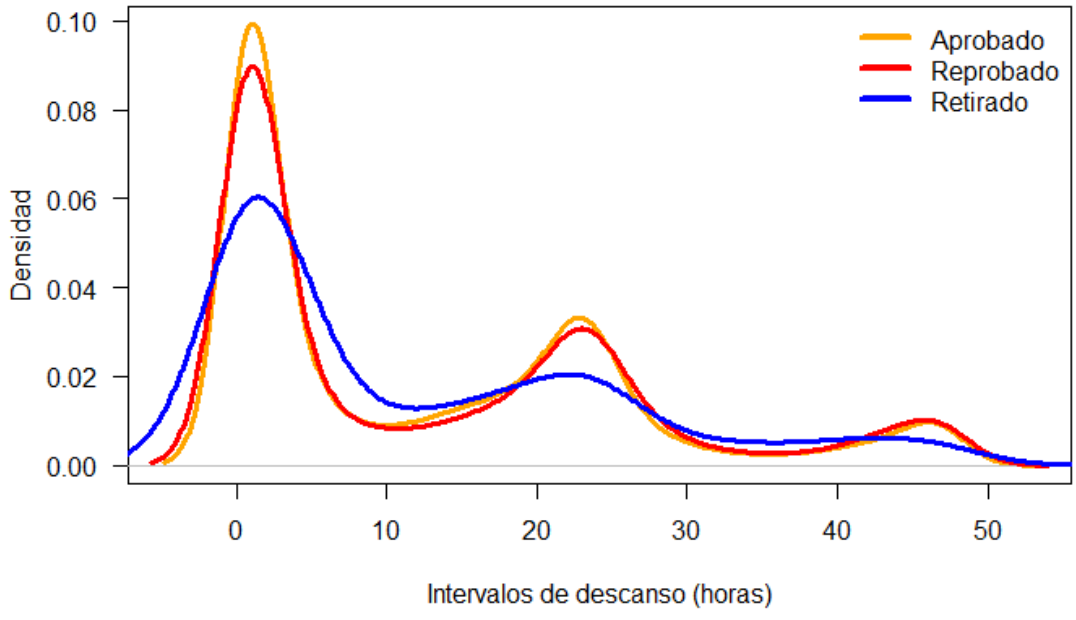

Figura 7.15: Función de distribución de descanso esperado para usuarios certificados.

Función de distribución de procrastinación ordinaria para usuarios certificados

Los intervalos de procrastinación ordinaria para los usuarios certificados se presentan en la figura 7.16. Se puede observar que para los tres grupos de usuarios la curva tiene una inclinación hacia la izquierda, es decir, los tiempos menores entre 48 y 336 horas son los más comunes. Ahora bien, las curvas para los usuarios reprobados y retirados se encuentran levemente hacia la derecha con respecto a la curva de los usuarios aprobados. Esto, a breves rasgos significa que, para periodos de procrastinación ordinaria, los tiempos de los usuarios reprobados y retirados son levemente mayores en comparación con el de los usuarios aprobados.

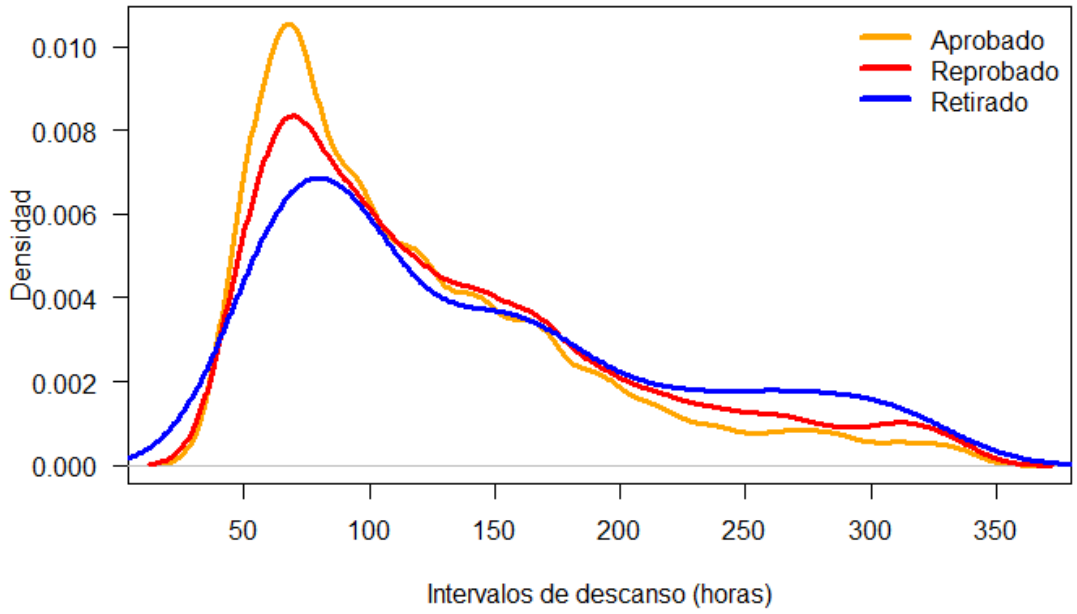

Figura 7.16: Función de distribución de procrastinación ordinaria para usuarios certificados.

Función de distribución de procrastinación extraordinaria para usuarios certificados

La procrastinación extraordinaria refleja periodos de inactividad entre sesiones más grandes que la duración del curso. Basándose en la figura 7.17, se puede observar que los tres grupos de estudiantes certificados presentan este comportamiento en periodos de tiempo que se aproximan a las 500 horas entre sesiones, puesto que el pico de sus curvas se encuentra hacia la izquierda. Pero, la curva de los usuarios retirados tiende a estar más elevada conforme esta se dirige hacia la derecha. Este comportamiento contrasta con lo deducido en la respuesta 2 de la sección anterior (R2), que los estudiantes etiquetados como retirados pagaron por el curso para tener acceso permanente a su material.

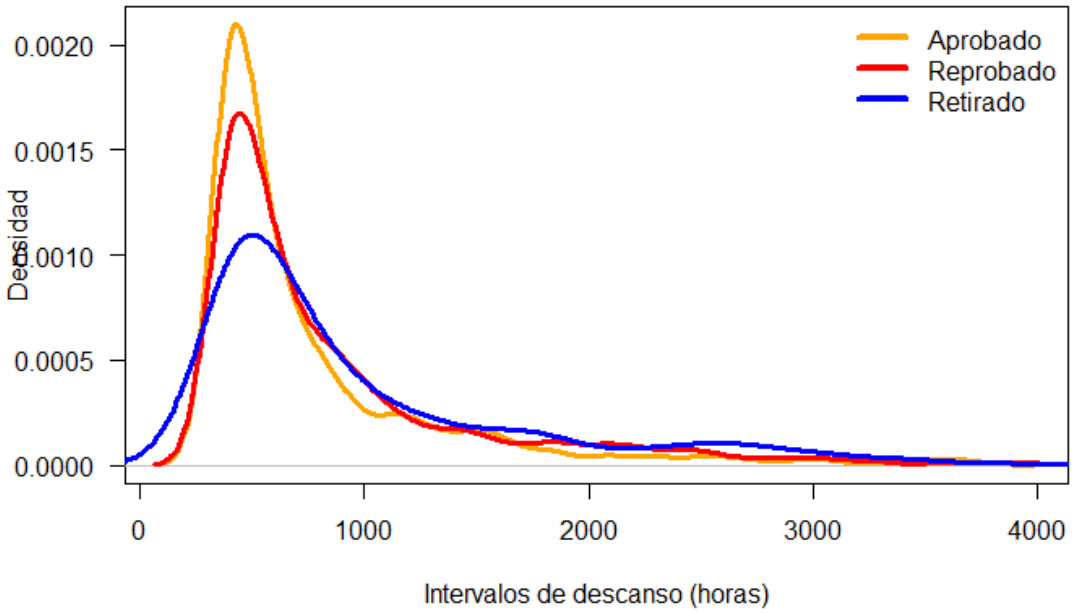

Figura 7.17: Función de distribución de procrastinación extraordinaria para usuarios certificados.

CAPÍTULO 8
Conclusiones y líneas de trabajo futuro

Las conclusiones generadas durante la realización de este trabajo de fin de máster son presentadas en este capítulo. Estas son desglosadas a detalle en la sección 8.1. Por otra parte, las líneas de trabajo futuro son expuestas en la sección 8.2.

8.1 Conclusiones

Los cursos MOOC son herramientas que han revolucionado el aprendizaje. Estas herramientas permiten a sus usuarios acceder a su contenido desde cualquier sitio, requiriendo solamente de una conexión a internet. Pero, sus bajas tasas de finalización han motivado a investigadores a desarrollar varios estudios en búsqueda del porqué de esta situación.

En este trabajo de fin de máster se realizó un caso de estudio que buscaba determinar variaciones en el comportamiento de los usuarios de un curso MOOC, pero, principalmente enfocándose en sí comportamientos contraproducentes, como la procrastinación, tienen consecuencias en el resultado obtenido por los usuarios, bajo esta premisa se obtuvieron las siguientes conclusiones:

1. Es necesario plantear estrategias de retención de usuarios efectivas para un curso MOOC. Esto con el fin de reducir la alta tasa de abandono que se presenta a partir de la segunda semana.
2. La falta de motivación de los usuarios no certificados que abandonan el curso se encuentra vinculada a las limitaciones que ofrece este en su versión gratuita.
3. En una sesión de trabajo, los usuarios que aprobaron el curso utilizan de mejor forma el material disponible.
4. Existe una mejor gestión de actividades por parte de los usuarios que aprobaron el curso, con respecto a los que lo reprobaron.
5. El material interactivo despierta el interés de los usuarios, ya sea actividades de evaluación o actividades de pronunciación. Queda en evidencia que este material tiende a provocar que el usuario mantenga su atención en el contenido.
6. Los audios dentro de un curso provocan desinterés, puesto que, existen sesiones de estudio que solo se enfocan en audios, audios cuyo tiempo máximo en conjunto no supera los dos minutos de reproducción.

7. Los usuarios tienden a procrastinar en una sesión de estudio cuando el material revisado es de poca o nula interacción. Materiales como los audios provocan desinterés y por lo tanto pérdida de tiempo dentro de una sesión de estudio, mientras que, otro material con poca interacción, como los videos, no tienen este mismo efecto. Basándose en lo planteado se puede inferir que material que involucra solamente uno de los cinco sentidos no llama la atención de los usuarios.

8. La procrastinación intra sesiones tiene relación con el éxito o el fracaso del usuario en el curso. Basándose en los resultados obtenidos, se puede observar que los usuarios que aprobaron tienen menos casos de procrastinación intra sesiones con respecto a quienes reprobaron. Los usuarios que aprobaron dirigen su atención a los videos en cuanto pierden la concentración, mientras que, los que reprobaron tienden a navegar entre el material sin volver a prestar atención al contenido.

9. La procrastinación entre sesiones no influye de gran forma en los resultados del curso. Puesto que, al analizar la distribución de los usuarios en las curvas de tiempo para este comportamiento, se pudo observar que su distribución es prácticamente igual.

Finalmente, como conclusión general a este caso de estudio se puede decir que la procrastinación tiene relación con el éxito o el fracaso del curso, pero, no es el único factor relevante. Puesto que, el tipo de material analizado y la forma que se recorre el contenido durante las sesiones de estudio también se encuentran relacionadas con el resultado final.

8.2 Líneas de trabajo futuro

A partir de este trabajo de fin de máster surgen diversas líneas de trabajo futuro, entre las cuales vale la pena resaltar:

1. Continuar con el análisis de los datos generados por usuarios de un curso MOOC desde otros enfoques. Buscando identificar detalladamente el comportamiento y solventar las falencias que puedan sufrir los usuarios a tiempo.

2. La procrastinación en cuanto a eventos evidencia los puntos en los que el usuario pierde el interés en el material, pero, un registro de eventos detallado basándose en la interacción con el contenido multimedia (reproducir, pausar, regresar) permitiría optimizar el uso de este tipo de material en los MOOC.

3. Diseñar y realizar estudios híbridos, que utilicen la minería de procesos de la mano con otras ciencias de la computación. De tal forma que, las ventajas de estas lleguen a complementarse, obteniendo así mejores resultados.

Bibliografía

[1] Y. Aljaraideh, «Massive Open Online Learning (MOOC) benefits and challenges: A case study in Jordanian context.,» *International Journal of Instruction*, vol. 12, n.º 4, págs. 65-78, 2019.

[2] Y. Jia, Z. Song, X. Bai y W. Xu, «Towards economic models for MOOC pricing strategy design,» en *International Conference on Database Systems for Advanced Applications*, Springer, 2017, págs. 387-398.

[3] *MOOCs Market to Grow By $ 17.70 bn During 2020-2024 | Industry Analysis, Market Trends, Market Growth, Opportunities, and Forecast | Technavio*, en, nov. de 2020. dirección: https://www.businesswire.com/news/home/20201117005958/en/MOOCs-Market-to-Grow-By-17.70-bn-During-2020-2024-Industry-Analysis-Market-Trends-Market-Growth-Opportunities-and-Forecast-Technavio (visitado 29-07-2021).

[4] B. Kitchenham, «Procedures for performing systematic reviews,» *Keele, UK, Keele University*, vol. 33, n.º 2004, págs. 1-26, 2004.

[5] M. N. S. RIZVI, «A Systematic Overview on Data Mining: concepts and techniques,» *International Journal of Research in Computer & Information Technology (IJRCIT) Vol*, vol. 1,

[6] F. Gorunescu, *Data Mining: Concepts, models and techniques*. Springer Science & Business Media, 2011, vol. 12.

[7] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[8] A. K. Pujari, *Data mining techniques*. Universities press, 2001.

[9] G. Kesavaraj y S. Sukumaran, «A study on classification techniques in data mining,» en *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, IEEE, 2013, págs. 1-7.

[10] D. L. Olson y D. Delen, *Advanced data mining techniques*. Springer Science & Business Media, 2008.

[11] Mike Chapple, *Use Regression to Predict Where You Might Live*, en. dirección: https://www.lifewire.com/regression-1019655 (visitado 29-07-2021).

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras y A. Vladu, «Towards deep learning models resistant to adversarial attacks,» *arXiv preprint arXiv:1706.06083*, 2017.

[13] C. dos Santos Garcia, A. Meincheim, E. R. F. Junior, M. R. Dallagassa, D. M. V. Sato, D. R. Carvalho, E. A. P. Santos y E. E. Scalabrin, «Process mining techniques and applications–a systematic mapping study,» *Expert Systems with Applications*, vol. 133, págs. 260-295, 2019.

[14] A. Bogarín, R. Cerezo y C. Romero, «A survey on educational process mining,» *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, n.º 1, e1230, 2018.

[15] W. M. van der Aalst y H. Verbeek, «Process discovery and conformance checking using passages,» *Fundamenta Informaticae*, vol. 131, n.º 1, págs. 103-138, 2014.

[16] M. De Leoni, W. M. Van Der Aalst y B. F. Van Dongen, «Data-and resource-aware conformance checking of business processes,» en *International Conference on Business Information Systems*, Springer, 2012, págs. 48-59.

[17] F. J. P. Hidalgo, C. A. H. Abril y col., «MOOCs: Origins, concept and didactic applications: A systematic review of the literature (2012–2019),» *Technology, Knowledge and Learning*, vol. 25, n.º 4, págs. 853-879, 2020.

[18] J. Kennedy, «Characteristics of massive open online courses (MOOCs): A research review, 2009-2012.,» *Journal of Interactive Online Learning*, vol. 13, n.º 1, 2014.

[19] L. A. Atiaja y R. Proenza, «The MOOCs: origin, characterization, principal problems and challenges in Higher Education,» *Journal of e-Learning and Knowledge Society*, vol. 12, n.º 1, 2016.

[20] N. Sharma, I. Doherty y D. Harbutt, «MOOCs and SMOCs: changing the face of medical education?» *Perspectives on medical education*, vol. 3, n.º 6, págs. 508-509, 2014.

[21] L. Song y J. R. Hill, «A conceptual model for understanding self-directed learning in online environments,» *Journal of Interactive Online Learning*, vol. 6, n.º 1, págs. 27-42, 2007.

[22] K. Mrhar, O. Douimi y M. Abik, «A Dropout Predictor System in MOOCs Based on Neural Networks,» *Journal of Automation, Mobile Robotics and Intelligent Systems*, págs. 72-80, 2021.

[23] J. J. Maldonado, R. Palta, J. Vázquez, J. L. Bermeo, M. Pérez-Sanagustín y J. Munoz-Gama, «Exploring differences in how learners navigate in MOOCs based on self-regulated learning and learning styles: A process mining approach,» en *2016 XLII Latin American Computing Conference (CLEI)*, IEEE, 2016, págs. 1-12.

[24] W. Matcha, D. Gašević, J. Jovanović, N. A. Uzir, C. W. Oliver, A. Murray y D. Gasevic, «Analytics of learning strategies: the association with the personality traits,» en *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 2020, págs. 151-160.

[25] J. Maldonado-Mahauad, M. Pérez-Sanagustín, P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino y C. Delgado-Kloos, «Predicting learners' success in a self-paced MOOC through sequence patterns of self-regulated learning,» en *European conference on technology enhanced learning*, Springer, 2018, págs. 355-369.

[26] J. Maldonado-Mahauad, M. Pérez-Sanagustín, R. F. Kizilcec, N. Morales y J. Munoz-Gama, «Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses,» *Computers in Human Behavior*, vol. 80, págs. 179-196, 2018.

[27] S. Li e Y. Zhang, «A Cluster Study on MOOC Students' Participation Patterns: A Case Study of a Chinese MOOC,» en *2018 Seventh International Conference of Educational Innovation through Technology (EITT)*, IEEE, 2018, págs. 184-188.

[28] P. Arpasat, P. Porouhan y W. Premchaiswadi, «Improvement of call center customer service in a thai bank using disco fuzzy mining algorithm,» en *2015 13th International Conference on ICT and Knowledge Engineering (ICT & Knowledge Engineering 2015)*, IEEE, 2015, págs. 90-96.

[29] M. L. Van Eck, X. Lu, S. J. Leemans y W. M. Van Der Aalst, «PM2: a process mining project methodology,» en *International Conference on Advanced Information Systems Engineering*, Springer, 2015, págs. 297-313.

[30] B. K. Chan, «Data analysis using R programming,» en *Biostatistics for Human Genetic Epidemiology*, Springer, 2018, págs. 47-122.

[31] A. Janes, F. M. Maggi, A. Marrella y M. Montali, «From Zero to Hero: A Process Mining Tutorial,» en *International Conference on Product-Focused Software Process Improvement*, Springer, 2017, págs. 625-629.

[32] G. Ongo y G. P. Kusuma, «Hybrid database system of MySQL and MongoDB in web application development,» en *2018 International Conference on Information Management and Technology (ICIMTech)*, IEEE, 2018, págs. 256-260.

[33] *2. Open edX Architecture — Open edX Developer's Guide documentation*. dirección: https://edx.readthedocs.io/projects/edx-developer-guide/en/latest/architecture.html (visitado 02-09-2021).

[34] Edx, *User Info and Learner Progress Data*. dirección: https://github.com/edx/edx-documentation/blob/master/en_us/data/source/internal_data_formats/sql_schema.rst.

[35] M. Vitiello, S. Walk, D. Helic, V. Chang y C. Guetl, «User Behavioral Patterns and Early Dropouts Detection: Improved Users Profiling through Analysis of Successive Offering of MOOC.,» *J. Univers. Comput. Sci.*, vol. 24, n.º 8, págs. 1131-1150, 2018.

[36] C.-H. Yu, J. Wu y A.-C. Liu, «Predicting learning outcomes with MOOC clickstreams,» *Education sciences*, vol. 9, n.º 2, pág. 104, 2019.

[37] Z. Ren, H. Rangwala y A. Johri, «Predicting performance on MOOC assessments using multi-regression models,» *arXiv preprint arXiv:1605.02269*, 2016.

[38] P. G. de Barba, D. Malekian, E. A. Oliveira, J. Bailey, T. Ryan y G. Kennedy, «The importance and meaning of session behaviour in a MOOC,» *Computers & Education*, vol. 146, pág. 103 772, 2020.

[39] I. Khan, M. Schommer-Aikins y N. Saeed, «Cognitive Flexibility, Procrastination, and Need for Closure Predict Online Self-Directed Learning Among Pakistani Virtual University Students,» *International Journal of Distance Education and E-Learning*, vol. 6, n.º 2, págs. 31-41, 2021.

[40] C. W. Günther y A. Rozinat, «Disco: Discover Your Processes.,» *BPM (Demos)*, vol. 940, págs. 40-44, 2012.