PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE HÁBITAT, INFRAESTRUCTURA Y CREATIVIDAD

MAESTRÍA EN SISTEMAS DE INFORMACIÓN MENCIÓN CIENCIA DE DATOS

PROYECTO DE TITULACIÓN

Tema:

Identificación de patrones de interacción en cursos MOOC para la predicción del rendimiento académico

Autor:

Toledo Illescas María Belén

Director:

Roa Marín Henry Nelson, PhD.

Quito DM, marzo 2025

Dedicatoria

A Dios, por darme la vida y la salud necesarias para cumplir mis objetivos.

A mi madre, Carmen, por su sabiduría, paciencia y amor con el que siempre me ha guiado.

A mi padre, Martin, por sus sacrificios y esfuerzos para brindarme lo mejor en cada etapa de mi vida.

A mi hermano, José, por ser mi ejemplo a seguir, mostrándome siempre que con esfuerzo puedo lograr todo lo que me propongo.

A mi hermana, Carmita, por motivar a ser mejor cada día y recordarme que nunca me debo rendir.

A mi abuelita, María, por su amor incondicional.

A mis amigos, que siempre han confiado en mí.

Agradecimiento

A Dios, por la vida, la salud y por bendecirme con una familia maravillosa, cuyo amor y apoyo han sido fundamentales en mi camino.

A mis padres, Martín y Carmen, por su incondicional, por ser mi fortaleza en los momentos difíciles y por enseñarme el valor del esfuerzo y la perseverancia.

Al PhD. Henry Roa, por su valioso tiempo, su guía y el apoyo brindado en la elaboración de este trabajo de titulación.

A mis docentes, por compartir su conocimiento inspirándonos a seguir aprendiendo y creciendo profesionalmente.

A mis amigos, por su confianza, ánimo y compañía en este proceso, haciéndolo más llevadero y significativo.

Resumen

Los Cursos Masivos Abiertos en Línea (MOOCs) han transformado el acceso a la educación, ofreciendo oportunidades de aprendizaje flexible y a gran escala, sin embargo, enfrentan el desafío de altas tasas de deserción y baja finalización. Este proyecto titulado "Identificación de patrones de interacción en cursos MOOC para la predicción del rendimiento académico" tiene como objetivo desarrollar un modelo predictivo que permita anticipar el rendimiento académico de los estudiantes en MOOCs mediante el análisis de la interacción y participación de los usuarios. Se analizaron las interacciones más frecuentes durante las sesiones de estudio, identificando patrones de aprendizaje diferenciados entre los estudiantes que aprobaron y aquellos que no completaron el curso. A través de herramientas de minería de procesos para comprender sus comportamientos y se evaluaron tres modelos de clasificación: Regresión logística, Máquinas de Soporte Vectorial (SVM) y Árboles de decisión, los resultados demostraron que el modelo de árbol de decisión alcanzó el mejor desempeño (accuracy = 90.86%, F1-score = 90.93%), logrando un equilibrio óptimo entre precisión y sensibilidad, se identificó que los estudiantes con mayor participación en video-lecturas, actividades suplementarias y evaluaciones tienen mayores probabilidades de aprobar el curso, mientras que aquellos con interacciones limitadas o desorganizadas presentan un mayor riesgo de abandono. Se concluye que los modelos predictivos pueden ser herramientas clave para identificar a estudiantes en riesgo y permitir intervenciones tempranas. Como trabajo futuro, se propone la integración del modelo en entornos educativos reales para generar alertas tempranas y recomendaciones en tiempo real, permitiendo a los instructores identificar a estudiantes en riesgo y brindarles apoyo oportuno.

Tabla de Contenidos

Contenido

Índice de tablas

Índice de figuras

1. Introducción

1.1. Justificación

La educación masiva y abierta ha experimentado un crecimiento exponencial en la última década gracias a las plataformas de Cursos Masivos Abiertos en Línea (MOOCs). Estas plataformas ofrecen una oportunidad única para democratizar el acceso a la educación, pero también presentan desafíos en términos de retención y finalización. Según García-Peñalvo et al. (2014), las tasas de finalización en los MOOCs son notablemente bajas, lo que plantea interrogantes sobre la efectividad de estos entornos de aprendizaje. A pesar de la creciente popularidad de los MOOCs, existe una brecha significativa entre la promesa de los cursos en línea y los resultados reales obtenidos por los estudiantes. Cáceres et al. (2020) destacan la necesidad de desarrollar estrategias basadas en analíticas de aprendizaje para mejorar el rendimiento estudiantil y la personalización del proceso educativo en estas plataformas.
El análisis de datos de interacción y comportamiento estudiantil dentro de los MOOCs, permite no solo entender mejor cómo los estudiantes interactúan con los recursos, sino también predecir su rendimiento académico. En este sentido, la presente investigación pretende desarrollar un modelo predictivo que anticipe el rendimiento de los estudiantes en base a patrones de interacción, brindando a las instituciones educativas una herramienta clave para realizar intervenciones tempranas y personalizadas. Este enfoque es crucial para mejorar las tasas de retención y finalización en MOOCs, permitiendo un uso más eficiente de los recursos educativos.

1.2. Formulación del problema

El auge de los MOOCs ha revolucionado el acceso a la educación a nivel global. Sin embargo, uno de los problemas más críticos es la baja tasa de finalización de estos cursos. Como destacan Gómez (2022) y Zapata-Ros (2013), solo una pequeña proporción de estudiantes que inician un MOOC lo completan con éxito. Este fenómeno plantea la necesidad de analizar los factores que influyen en el rendimiento académico de los estudiantes dentro de estas plataformas. En particular, la falta de interacción sostenida con los recursos educativos y la falta de personalización en los enfoques pedagógicos podrían ser factores determinantes del bajo rendimiento.

El problema radica en la dificultad de identificar y predecir de manera precisa el rendimiento académico de los estudiantes en un MOOC a partir de los patrones de interacción y participación. Mientras que algunos estudiantes interactúan activamente con los foros, videos y actividades del curso, otros muestran conductas pasivas o abandonan rápidamente. La investigación de Kloos et al. (2016) sugiere que los patrones de interacción pueden ser utilizados para predecir el éxito o fracaso de los estudiantes, pero se necesitan más estudios que exploren cómo estas interacciones varían entre diferentes tipos de estudiantes y cómo se pueden utilizar para intervenciones personalizadas.

1.3. Contextualización del tema

Este proyecto de investigación se centra en la identificación y análisis de los patrones de interacción en un curso MOOC para desarrollar modelos predictivos que anticipen el rendimiento académico de los estudiantes. El uso de las analíticas de aprendizaje en la educación superior, y específicamente en los MOOCs, ha ganado relevancia en los últimos años, como lo señalan Cáceres et al. (2020) y Ruipérez-Valiente (2020). Este campo emergente combina la ciencia de datos y la educación para analizar grandes volúmenes de datos generados por la interacción de los estudiantes con los recursos y actividades de un curso en línea.

En el nivel macro, los MOOCs presentan un desafío global en términos de acceso y éxito educativo, donde millones de estudiantes pueden inscribirse, pero pocos completan los cursos. A nivel meso, las plataformas educativas como edX y Coursera generan enormes cantidades de datos sobre las interacciones de los estudiantes, que pueden ser utilizados para mejorar el diseño del curso y las estrategias pedagógicas. A nivel micro, cada estudiante interactúa de manera única con los materiales del curso, y estas interacciones pueden revelar patrones que predicen su éxito o fracaso, como menciona Gómez (2022). Por este motivo se propone analizar esos patrones de interacción desde múltiples perspectivas (macro y micro) para construir modelos predictivos basados en la identificación de secuencias de aprendizaje. De acuerdo a Kloos et al. (2016), la implementación de herramientas de analíticas de aprendizaje puede mejorar la retención y el rendimiento al ofrecer información personalizada y oportuna a los estudiantes y educadores.

1.4.Objetivos

1.4.1. Objetivo General

Desarrollar un modelo predictivo basado en la identificación de patrones de interacción y participación en cursos MOOC que permita anticipar el rendimiento académico de los estudiantes y mejorar las tasas de finalización y retención.

1.4.2. Objetivos Específicos

- Analizar las interacciones más frecuentes de los estudiantes en una sesión de estudio en un MOOC.
- Comparar las secuencias de aprendizaje entre los estudiantes que aprobaron y los que no aprobaron el curso.
- Desarrollar un modelo predictivo que anticipe el rendimiento académico en función de los patrones de interacción.

1.5.Preguntas investigación

OE1: Analizar las interacciones más frecuentes de los estudiantes en una sesión de estudio en un MOOC

P.I.1. ¿Cuáles son las interacciones más frecuentes de los estudiantes en una sesión de estudio en un MOOC?

P.I.2. ¿Qué patrones de aprendizaje se encontraron?

OE2: Comparar las secuencias de aprendizaje entre los estudiantes que aprobaron y los que no aprobaron el curso.

P.I.3. ¿Existen diferencias en las secuencias de aprendizaje entre los estudiantes que lograron aprobar el curso y aquellos que no lo hicieron?

OE3 Desarrollar un modelo predictivo que anticipe el rendimiento académico en función de los patrones de interacción.

P.I.4. ¿Cómo pueden los modelos predictivos, basados en patrones de interacción, identificar a los estudiantes en riesgo de bajo rendimiento en un curso MOOC?

1.6. Marco Conceptual y Marco Teórico

1.6.1. Revisión de Literatura

En la última década, los Cursos Masivos Abiertos en Línea (MOOCs) han revolucionado el acceso a la educación, proporcionando oportunidades de aprendizaje sin restricciones económicas ni geográficas. Plataformas como Coursera, edX y MiriadaX han facilitado la difusión de contenido académico de calidad, aunque estos cursos enfrentan retos importantes en términos de retención y finalización. Según Ruipérez-Valiente et al. (2020), la baja tasa de finalización de los MOOCs se debe, en parte, a la falta de interacción efectiva dentro de la plataforma. Por ello, la investigación académica ha centrado su atención en analizar los patrones de participación con el fin de mejorar la permanencia estudiantil.

Un aspecto clave en la investigación sobre MOOCs es el estudio de la organización del tiempo de los estudiantes y su influencia en el desempeño académico. Diversos estudios han examinado la relación entre la participación en foros, la entrega de tareas y el tiempo dedicado al curso. Wu (2021) encontró que una mayor actividad en foros se asocia con percepciones más positivas del curso y mayores tasas de finalización. Además, Wu y Li (2020) analizaron el impacto del "boca a boca electrónico" (eWOM), concluyendo que las recomendaciones y discusiones en línea influyen en la inscripción y permanencia de los estudiantes en estos cursos.

Bianchi et al. (2022) argumentan que, si bien los MOOCs son importantes para la educación superior, se debe trabajar en mejorar la permanencia de los estudiantes en estos cursos. Implementar enfoques basados en la identificación de patrones de aprendizaje e interacción resulta fundamental para afrontar estos desafíos y optimizar la experiencia de los participantes.

La predicción del rendimiento académico a través del análisis de patrones de interacción constituye una línea de investigación importante. Investigaciones que se han realizado con anterioridad, han utilizado técnicas de machine learning para poder diferenciar a los estudiantes que completan los cursos y aquellos que los abandonan. Geigle y Zhai (2017) aplicaron un modelo oculto de Márkov (HMM) para evaluar la relación entre patrones de

participación y el éxito en estos cursos distinguiendo diferencias significativas entre estudiantes aprobados y no aprobados, encontrando que aquellos con mejor desempeño dedicaban más tiempo a la realización de pruebas y participaban activamente en foros.

Diversos factores han sido documentados en relación con el abandono de los MOOCs. Investigaciones como las de Bernal-González (2015) ha identificado causas como la falta de tiempo, conocimientos previos insuficientes, diseño instruccional deficiente y escasa interacción con docentes y compañeros. Además, la falta de autorregulación del aprendizaje es un factor crítico en la retención estudiantil.

Desde la perspectiva de la minería de procesos y analítica del aprendizaje, el estudio de Maldonado-Mahauad et al. (2018) identificó distintos tipos de estudiantes en MOOCs, clasificándolos acorde a sus patrones de interacción, se diferenciaron aquellos que siguen un aprendizaje estructurado "Sampling Learners", de quienes interactúan de manera esporádica "Targeting Learners". La implementación de metodologías basadas en datos ha permitido identificar factores clave que inciden en el rendimiento académico, lo que refuerza la necesidad de intervenciones tempranas para mejorar la experiencia de aprendizaje en plataformas digitales

Trigwell & Prosser (1991) exploraron dos enfoques estrategias de aprendizaje uno basado en aptitudes el cual se apoya en autoevaluaciones de los estudiantes sobre sus métodos de aprendizaje y otro basado en el proceso el cual analiza eventos derivados de la interacción con la plataforma.

En analíticas del aprendizaje Jovanović et al. (2017) aplicaron minería de secuencias y aprendizaje automático no supervisado para extraer estrategias de aprendizaje en entornos virtuales(Moodle) y encontraron patrones como el seguimiento estándar del curso, el uso inadecuado de la retroalimentación y la multitarea durante evaluaciones. De manera similar, Finchman et al. (2018) emplearon modelos de Markov para agrupar patrones de estudio en estrategias específicas. Mukala et al. (2015) analizaron a más de 40,000 estudiantes en un MOOC, concluyendo que aquellos con calificaciones más altas presentaban patrones de interacción más estructurados, mientras que los de menor desempeño mostraban comportamientos más irregulares.

1.6.2. Marco Conceptual

MOOCs

Los Cursos Masivos Abiertos en Línea (MOOCs) surgieron a finales de la década de 2000 como una nueva modalidad educativa destinada a democratizar el acceso al conocimiento. Inicialmente desarrollados por universidades de prestigio, como el MIT y Stanford, estos cursos permitieron que personas de todo el mundo accedieran a formación de alta calidad de manera gratuita o a bajo costo. Según García-Peñalvo et al. (2014), los MOOCs representan una innovación disruptiva en la educación superior, ya que permiten la participación masiva de estudiantes con diversas formaciones y habilidades, independientemente de su ubicación geográfica. La flexibilidad y el acceso ilimitado son características fundamentales de los MOOCs, que les han permitido atraer a millones de estudiantes. Sin embargo, uno de los desafíos persistentes es la baja tasa de finalización, con estudios que sugieren que menos del 10% de los inscritos completan los cursos, un fenómeno atribuido a la falta de interacción directa con los instructores y el bajo compromiso de los estudiantes (García-Peñalvo et al., 2014).

Los MOOCs han evolucionado significativamente desde su creación debido a que inicialmente fueron concebidos como una alternativa educativa masiva, su estructura pedagógica ha ido adaptándose a las necesidades de diferentes tipos de estudiantes. Estos cursos suelen dividirse en módulos semanales, con contenido en video, lecturas complementarias, foros de discusión y evaluaciones automáticas. Pese a sus ventajas, uno de los retos principales es mantener a los estudiantes comprometidos durante todo el curso. La falta de interacción en tiempo real, la desmotivación y el desinterés son algunas de las barreras que impiden que muchos alumnos finalicen los MOOCs (Wu & Li, 2020).

Coursera

Coursera es una plataforma de educación en línea basada en el modelo MOOC (Massive Open Online Courses). Este modelo educativo prioriza el aprendizaje y el acceso al conocimiento por encima de la obtención de certificaciones. La plataforma fue desarrollada por un grupo de profesionales de la Universidad de Stanford en octubre de 2011. Sus

contenidos son elaborados por diversas instituciones académicas de prestigio, entre las que destacan Stanford University, University of Michigan, Johns Hopkins University y Duke University. Además, Coursera ofrece cursos en múltiples idiomas, incluyendo inglés, español, portugués, chino, francés y ruso (Martin & Ramírez, 2016).

La plataforma proporciona cursos en áreas de actualidad, como inteligencia artificial, aprendizaje automático (Machine Learning) y aprendizaje profundo (Deep Learning). Aunque el acceso a los cursos es gratuito en muchos casos, los usuarios tienen la opción de pagar para obtener un certificado que valide los conocimientos adquiridos. Los cursos están organizados en secciones estructuradas por lecciones o semanas, donde se incluyen video-lecciones y evaluaciones. En cuanto a los cronogramas y sesiones del curso, estos establecen los períodos en los que los participantes pueden realizar las actividades de manera conjunta (Coursera, 2019).

La estructura de un curso en Coursera generalmente sigue el siguiente esquema:

- **Información del curso:** Presenta los objetivos, competencias a desarrollar y los profesionales e instituciones responsables del curso.
- **Vista preliminar del curso:** Permite visualizar la planificación y organización del contenido por semanas. Al seleccionar una semana específica, se accede a los siguientes materiales:
    - **Videos:** Lecciones grabadas por docentes con explicaciones detalladas. Un tema puede contar con varias video-lecciones.
    - **Foros:** Espacios de discusión para la interacción entre estudiantes, docentes y el equipo de apoyo técnico.
    - **Cuestionarios:** Evaluaciones automatizadas diseñadas para medir el nivel de comprensión del alumnado.
    - **Tareas:** Actividades que pueden incluir ensayos revisados por pares, ejercicios de programación y problemas matemáticos (Martin & Ramírez, 2016).

1.6.3. Marco Teórico

Analítica de aprendizaje (Learning Analytics)

Las analíticas de aprendizaje consisten en el conjunto de herramientas y técnicas destinadas a recopilar, analizar y utilizar datos educativos para mejorar el proceso de enseñanza-aprendizaje. Su objetivo principal es ayudar a los educadores y administradores a entender cómo los estudiantes interactúan con los cursos y tomar decisiones informadas para mejorar su experiencia educativa. Dentro de los MOOCs, la analítica de aprendizaje ha ganado especial relevancia, ya que permite analizar grandes volúmenes de datos generados por miles o millones de estudiantes (Ruipérez-Valiente, 2020). Estas herramientas ayudan a identificar patrones de comportamiento, predecir el rendimiento académico y planificar intervenciones que puedan mejorar las tasas de finalización.

El proceso de analítica de aprendizaje comienza con la recolección de datos a través de plataformas educativas que registran las interacciones de los estudiantes. A continuación, se procesan y visualizan estos datos para identificar patrones relevantes, como la frecuencia de acceso a los materiales del curso, la participación en los foros y la finalización de tareas. Según Ruipérez-Valiente et al. (2020), uno de los principales beneficios de la analítica de aprendizaje es su capacidad para generar modelos predictivos que anticipen el rendimiento de los estudiantes, lo que permite a los educadores diseñar intervenciones personalizadas. Cáceres et al. (2020) destacan que estos modelos que se crean se basan en datos históricos, como la frecuencia de acceso a los contenidos y el tiempo dedicado a las actividades, lo que permite reconocer a los estudiantes con alta probabilidad de deserción.

Modelos predictivos en la educación

Los modelos predictivos en la educación son herramientas analíticas que utilizan datos históricos y patrones de comportamiento para anticipar los resultados academicos de los estudiantes. Estos modelos se han aplicado en diversos contextos educativos, incluyendo los MOOCs, donde la tasa de abandono es significativamente alta. Según Gómez (2022), los modelos predictivos en educación permiten identificar a los estudiantes que están en riesgo de abandonar el curso antes de que suceda, lo que facilita la implementación de

estrategias de intervención para mejorar su desempeño y motivación. Los modelos predictivos utilizan una amplia gama de datos, como el tiempo de conexión, el número de tareas completadas, la participación en foros y el acceso a los materiales del curso. Estos datos permiten predecir si un estudiante logrará completar el curso o si abandonará antes de tiempo. Además, estos modelos también pueden identificar patrones de comportamiento que son indicativos de éxito académico, lo que permite a los educadores diseñar experiencias de aprendizaje más efectivas. En estudios recientes, se ha demostrado que los modelos predictivos pueden ser particularmente útiles para personalizar el aprendizaje y proporcionar retroalimentación en tiempo real a los estudiantes (Wu, 2021).

Patrones de interacción estudiantil

Los patrones de interacción y participación estudiantil hacen referencia a las secuencias de acciones y comportamientos que los estudiantes realizan en un entorno de aprendizaje en línea. Estos patrones son un indicador clave del compromiso y la motivación de los estudiantes, y están estrechamente relacionados con su rendimiento académico. En el contexto de los MOOCs, las interacciones incluyen el tiempo dedicado a ver videos, la participación en foros, la finalización de tareas y la revisión de materiales complementarios. Según Wu (2021), la identificación de estos patrones permite a los educadores entender mejor cómo los estudiantes se relacionan con el curso y qué factores influyen en su éxito o abandono.

La medición de los patrones de interacción se basa en la recopilación de datos de las plataformas educativas y mediante el análisis de estos datos, es posible identificar comportamientos comunes entre los estudiantes exitosos, como la frecuencia de acceso a los contenidos o la participación activa en foros de discusión. Los estudios han demostrado que los estudiantes que interactúan regularmente con el material del curso y participan en discusiones tienden a obtener mejores resultados académicos (García-Peñalvo et al., 2014).

Modelos de Predicción

Los modelos de predicción son herramientas fundamentales en la ciencia de datos, utilizados para anticipar resultados futuros basados en conjuntos de datos históricos estos

modelos emplean algoritmos de aprendizaje automático y técnicas estadísticas para identificar patrones y relaciones que permiten realizar predicciones precisas. Su aplicación en la actualidad es muy variada y los podemos encontrar desde la predicción del comportamiento de un consumidor hasta en el pronóstico de tendencias en el mercado financiero (AWS,2024).

La selección del modelo de predicción adecuado depende de la naturaleza de los datos y del problema a resolver por lo cual es crucial considerar factores como la precisión del modelo, su interpretabilidad y su capacidad para generalizar a nuevos datos.

Regresión Logística

La regresión logística es un modelo de clasificación binaria que estima la probabilidad de que un dato u observación se clasifique en una de dos categorías, modelando la relación entre las variables predictoras y la probabilidad del resultado a través de la función logística. A diferencia de la regresión lineal, que estima valores continuos, la regresión logística aplica una función logística para modelar la probabilidad de pertenencia a una categoría específica. Por ejemplo, es común su uso en la detección de spam, donde el modelo determina la probabilidad de que un correo electrónico sea no deseado (IBM, 2024).

Árboles de Decisión

Los árboles de decisión son modelos no paramétricos que representan decisiones y sus posibles consecuencias a través de una estructura de árbol jerárquica, en el que cada punto de decisión evalúa una característica, cada bifurcación muestra el resultado de esa evaluación y cada punto final ofrece una conclusión o clasificación. Su capacidad para manejar tanto variables categóricas como continuas los hace versátiles en diversas aplicaciones, como la segmentación de clientes y la evaluación de riesgos (Ciencia de Datos, 2020).

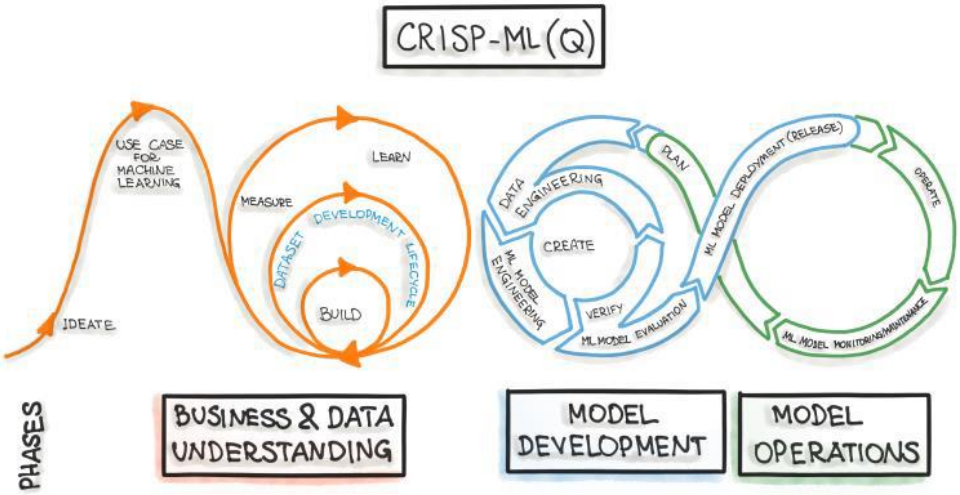
Máquinas de Vectores de Soporte (SVM)

SVM o Support Vector Machine, por sus siglas en inglés son algoritmos de aprendizaje supervisado utilizados para clasificación y regresión, aquí se busca encontrar el hiperplano que divide mejor los datos en categorías, procurando la mayor separación posible entre ellas. Son especialmente útiles en escenarios donde las clases no son linealmente separables, gracias al uso de funciones kernel que transforman los datos a espacios de mayor dimensión para facilitar la separación. Las SVM han demostrado una alta precisión en comparación con otros clasificadores, aplicándose en áreas como la detección de rostros y la clasificación de textos

CRISP-ML(Q)

Dentro del ámbito de la ciencia de datos, la metodología CRISP-ML(Q) se ha consolidado como una estructura de trabajo para la generación de sistemas que implementan aprendizaje automático (ML) con un enfoque riguroso en el aseguramiento de la calidad. Este modelo, una extensión adaptada del modelo CRISP-DM, facilita la estructuración de proyectos de ML desde la comprensión inicial del problema hasta el despliegue y mantenimiento del modelo (Ambler, 2020). CRISP-ML(Q) se compone de seis fases interconectadas: comprensión del negocio y los datos, ingeniería de datos, modelado, aseguramiento de la calidad, despliegue, y monitoreo y mantenimiento del modelo, como se ilustra en la **¡Error! No se encuentra el origen de la referencia.**.

Figura 1. *Fases del modelo CRISP-ML(Q). Fuente: ml-ops.org, 2024*



La adopción de CRISP-ML(Q) en proyectos de ciencia de datos se justifica por su capacidad para garantizar la calidad en cada etapa del ciclo de vida del modelo, minimizando riesgos como el sobreajuste y la falta de reproducibilidad. La fase inicial, "Comprensión del negocio y de los datos", es crucial para definir claramente los objetivos del modelo y verificar la disponibilidad y viabilidad de los datos. En la "Ingeniería de datos" se enfoca en la preparación y limpieza de los datos, un paso esencial para asegurar la calidad de los datos de entrada. En la fase de "Modelado", se seleccionan y entrenan los algoritmos de ML más adecuados, buscando un equilibrio entre rendimiento, interpretabilidad y robustez. Finalmente, las fases de "Despliegue" y "Monitoreo y mantenimiento" aseguran la correcta implementación y el rendimiento continuo del modelo. Para cumplir con los objetivos de la investigación, este marco metodológico proporcionará una estructura sólida para el análisis de los patrones de interacción y la predicción del rendimiento académico en MOOCs, permitiendo una aplicación sistemática y rigurosa de las técnicas de aprendizaje automático

SMOTE

Synthetic Minority Over-sampling Technique es una técnica utilizada para abordar el problema del desbalance de clases en conjuntos de datos mediante la generación de nuevas instancias sintéticas de una clase minoritaria en lugar de simplemente replicar las existentes. Para ello, SMOTE crea ejemplos intermedios interpolando entre los puntos de datos reales más cercanos en el espacio de características, lo que ayuda a mejorar el rendimiento de los modelos de aprendizaje automático al proporcionar una mejor representación de la clase minoritaria (Chawla et al., 2002). Esta técnica ha demostrado ser efectiva en la mejora de la generalización de los clasificadores, especialmente en problemas donde el desbalance afecta negativamente la capacidad predictiva del modelo.

Minería de Procesos

En la actualidad existen un sin número de sistemas implementados en las organizaciones, y la mayoría de estos se encargan de los procesos de negocio, almacenando información clave sobre actividades, eventos y tiempos relacionados con la ejecución de sus operaciones (Aguirre & Rincón, 2015). La minería de procesos permite analizar estos registros para obtener una visión más clara del funcionamiento real de los procesos y aplicar mejoras

basadas en datos, buscando descubrir patrones dentro de la ejecución de los procedimientos organizacionales. Para ello, se emplea un conjunto de datos que contiene los registros detallados de eventos (Vega, 2016). Mediante este enfoque, las empresas pueden:

1. Obtener una representación precisa del proceso real a partir de datos históricos.
2. Comparar el modelo derivado con los procedimientos documentados para evaluar su cumplimiento con normativas y políticas internas.
3. Evaluar la interacción de los colaboradores en los procesos para identificar posibles redundancias o ineficiencias.
4. Detectar cuellos de botella analizando el flujo de eventos y tiempos de ejecución.
5. Estimar la duración de procesos mediante modelos predictivos basados en técnicas de clasificación y árboles de decisión.
6. Identificar las variables clave que inciden en la eficiencia de los procesos y optimizar su rendimiento.

Por lo que Aguirre y Rincón (2015) establecen que la minería de procesos tiene como propósito fundamental la optimización de los procesos, por medio del análisis de los datos organizacionales, para eso se emplean diversas herramientas.

Herramienta para minería de procesos - DISCO

En la actualidad, existen diversas herramientas para la minería de procesos, entre ellas Disco, un software desarrollado por Fluxicon que permite descubrir, analizar y optimizar procesos empresariales a partir de registros de eventos. Su funcionalidad principal radica en la capacidad de importar y procesar grandes volúmenes de datos de manera eficiente. Para ello, Disco identifica automáticamente las marcas de tiempo en archivos CSV o Excel, asegurando un manejo ágil de la información. La configuración de las columnas incluye principalmente el ID del caso, las marcas de tiempo y los nombres de las actividades, aunque también pueden añadirse otros atributos relevantes para el análisis. Es importante destacar que los datos importados permanecen en modo de solo lectura, garantizando su integridad durante el procesamiento (Fluxicon, 2024).

Una vez importados los datos, el software genera un mapa visual del proceso, donde se representan las actividades y sus relaciones a través de una codificación de colores y distintos grosores de línea. Este enfoque facilita la identificación de patrones, la detección de bucles de procesamiento y el reconocimiento de ineficiencias en el flujo de trabajo. Disco se basa en el algoritmo Fuzzy Miner, pionero en la introducción de la "metáfora de mapa" en minería de procesos. Su diseño combina experiencia práctica con pruebas de usuario, logrando un equilibrio entre precisión y facilidad de uso. Gracias a esto, usuarios con conocimiento en el dominio, pero sin experiencia en minería de procesos, pueden operar el sistema eficientemente (Fluxicon, 2024).

Además del descubrimiento de procesos, Disco proporciona estadísticas detalladas sobre el número de casos y eventos registrados, ofreciendo métricas de frecuencia y rendimiento para cada actividad y recurso. Esta información permite evaluar la eficiencia del proceso y detectar posibles áreas de mejora. También permite el análisis de casos individuales, lo que resulta útil para identificar desviaciones, anomalías o incumplimientos de reglas de negocio. En este contexto, una variante se define como una secuencia específica de actividades, es decir, un camino desde el inicio hasta el final del proceso. Analizar estas variantes ayuda a detectar comportamientos atípicos dentro del flujo de trabajo (Fluxicon, 2024).

Para facilitar el análisis, el software ofrece opciones avanzadas de filtrado. Entre los principales filtros se encuentran aquellos que permiten seleccionar eventos por rangos de fechas, enfocarse en variantes específicas, filtrar casos según actividades de inicio y fin, o aplicar filtros basados en atributos de datos. Asimismo, se puede emplear el filtro Follower, útil para identificar patrones y verificar posibles infracciones de tareas. Estos mecanismos permiten explorar los procesos de forma interactiva y responder preguntas concretas sobre su funcionamiento (Fluxicon, 2024).

Otro aspecto clave del análisis en minería de procesos es el rendimiento. Disco calcula automáticamente la duración promedio de cada actividad y los tiempos de espera entre ellas a partir de las marcas de tiempo, facilitando la identificación de cuellos de botella y oportunidades de optimización (Fluxicon, 2024). Adicionalmente, la herramienta cuenta

con una función de animación, que permite visualizar la ejecución del proceso en tiempo real. Es importante no confundir la animación con la simulación, ya que esta última busca predecir escenarios futuros, mientras que la animación es una herramienta útil para comunicar los hallazgos a personas no expertas en análisis de procesos (Fluxicon, 2024).

1. Metodología de la investigación

La presente investigación sigue el modelo CRISP-ML en la fase de comprensión del negocio y datos, se analiza la estructura del curso y las fuentes de información disponibles, identificando las variables clave que influyen en el desempeño estudiantil, evaluando su calidad, completitud y posibles inconsistencias. En la fase de ingeniería de datos consiste en integrar múltiples archivos para construir un log estructurado de eventos, definiendo sesiones de usuario, marcas temporales, tipos y cantidades de actividades, se seleccionan las características más relevantes y se conforman el dataset específico que incluye solo información hasta el 50% del curso, permitiendo realizar predicciones tempranas. En el modelado, se entrenan modelos de aprendizaje automático para predecir la aprobación del curso, probando distintos algoritmos y evaluando su desempeño con métricas como precisión y determinando que modelo ofrece mejores resultados. En la fase de despliegue, se propone la integración del modelo a una aplicación que brinde retroalimentación temprana al usuario, facilitando intervenciones oportunas para mejorar la tasa de aprobación.

2.1. Comprensión del negocio y los datos

El curso "Electrones en acción" introduce a los estudiantes en la electrónica y el uso de Arduino, abarcando desde la elaboración de circuitos básicos hasta proyectos sencillos con dispositivos programables. Dado que el curso está orientado a estudiantes que cursan la etapa final de la educación secundaria y el inicio de la universidad, se promueve un aprendizaje basado en la práctica, donde los estudiantes participan activamente.

El curso está estructurado en cuatro módulos, cada uno de los cuales incluye diversas actividades, como video-lecturas y evaluaciones tanto formativas como sumativas. El contenido del curso abarca lecturas complementarias, material suplementario y cuestionarios (quizzes), siendo estos últimos los únicos de carácter obligatorio para aprobar el curso.

Para el análisis del curso, se consideró la participación de los 8,017 estudiantes registrados, quienes formaron parte del estudio.

La extracción de datos se realizó a partir de un conjunto de datos está compuesto por 86 archivos en formato CSV, cada uno acompañado de un archivo HTML que contiene información relevante sobre su contenido, como el nombre de la tabla, su descripción, las columnas con sus respectivas explicaciones y el código SQL para la creación de la tabla en una base de datos relacional (Figura 2). Durante este proceso, se identificaron las tablas necesarias para responder las preguntas de investigación.

***Figura 2**. Ejemplo del archivo HTML de la tabla o archivo assesment_questions, Fuente Coursera*

assessment_actions

Description

[No table description available]

Columns

Name	Description
assessment_action_id	No column description available
assessment_action_base_id	No column description available
assessment_id	No column description available
assessment_scope_id	No column description available
assessment_scope_type_id	No column description available
assessment_action_version	No column description available
assessment_action_ts	No column description available
assessment_action_start_ts	No column description available
guest_user_id	No column description available
ucchile_assessments_user_id	Encrypted Coursera user id for ucchile assessments data.

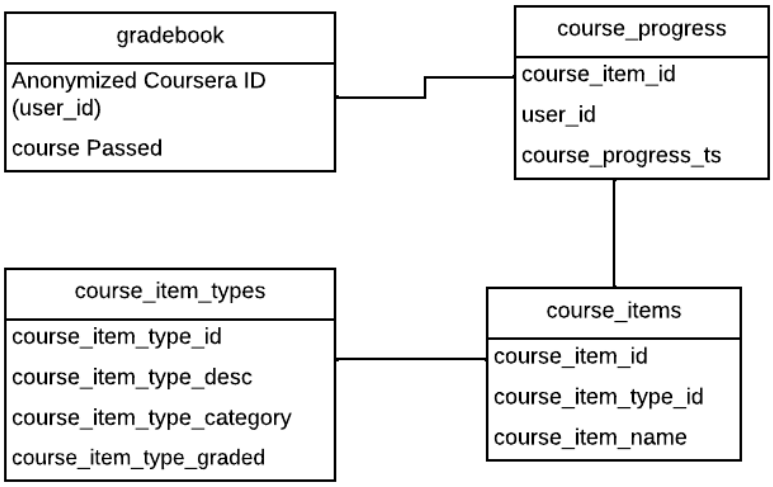
SQL create statement

```
CREATE TABLE assessment_actions (
    assessment_action_id VARCHAR(100)
    ,assessment_action_base_id VARCHAR(100)
    ,assessment_id VARCHAR(50)
    ,assessment_scope_id VARCHAR(50)
    ,assessment_scope_type_id INT4
    ,assessment_action_version INT4
    ,assessment_action_ts TIMESTAMP
    ,assessment_action_start_ts TIMESTAMP
    ,guest_user_id VARCHAR(50)
    ,ucchile_assessments_user_id VARCHAR(50) NOT NULL
);
```

Además, se contó con archivo adicional “gradebook” la cual contiene más de 40 variables algunas de ellas eran las calificaciones de las actividades sumativas dentro de cada módulo,

el timestamp donde se registra tanto la última rendición de la actividad como el timestamp de 12

Figura 3. *Modelo de datos simplificado. Fuente. Elaboración propia*



A continuación, se describen las tablas seleccionadas y las columnas fue elegidas en función a su relevancia para responder las preguntas de investigación, asegurando que la información extraída sea pertinente y significativa para el análisis.

- **gradebook:** Almacena los datos de las calificaciones sumativas de cada estudiante en las evaluaciones de los módulos(semanas), así como el estado de aprobación del curso.
    - **Anonymized Coursera ID (user_id):** El identificador de los usuarios dentro del curso MOOC, electrones en acción.
    - **course Passed:** Es un identificador el cual indica si el estudiante aprobo o reprobo el curso MOOC.
- **course_progress**: Almacena los datos relacionados con el avance de los usuarios en el curso, en la cual se incluyen las actividades calificadas y no calificadas, se registra el estado de avance del usuario, el cual puede tener dos estados, iniciado o completado (1,2 respectivamente), junto con su marca de temporal.
    - **course_item_id:** Columna que identifica un elemento individual dentro de un curso, es decir identifica la actividad que realiza X estudiante dentro del curso.

    - **user_id:** Código para cada usuario, de esta manera coursera maneja el anonimato de cada usuario.
    - **course_progress_ts:** La marca de tiempo para cuando el progreso de un alumno en una actividad(ítem) del curso ha cambiado.
- **course_items**: Almacena información relativa al nombre, id y tipo de cada actividad, y su función principal es permitir la conexión, a través de los identificadores, otras tablas que complementan el contexto del modelo desarrollado para el caso de estudio.
    - **course_item_id:** Columna que identifica un elemento individual dentro de un curso, es decir identifica la actividad que realiza X estudiante dentro del curso.
    - **course_item_type_id:** Dentro del curso MOOC hay muchos tipos diferentes de actividades(ítems) que componen un curso. Cada elemento recibe un item_type_id para facilitar su identificación.
    - **course_item_name:** El nombre de una actividad(ítem), como se ve en la vista del alumno del curso.
- **course_items_types**: Son las actividades que se pueden desarrollar dentro del curso MOOC, se integran al modelo mediante el identificador del tipo de actividad presente en la tabla “course_items”.
    - **course_item_type_id:** Dentro del curso MOOC hay diferentes tipos de actividades(ítems) que componen un curso. Cada elemento recibe un item_type_id para facilitar su identificación.
    - **course_item_type_desc:** Describe el tipo de actividad de una forma reducida.
    - **course_item_type_category:** Existen varias categorías diferentes en las que se encuentran las actividades del curso, por ejemplo: conferencias, complementarias/suplementarias (es decir, lectura), cuestionario, revisión por pares y programación. Dentro de cada una de estas categorías, existe la posibilidad de tener varios tipos de ítems, que permiten un manejo diferente de los ítems (por ejemplo, tareas formativas o sumativas y revisiones por pares abiertas o cerradas).

- **course_item_type_graded:** Una actividad puede ser calificada ("verdadero", lo que significa que se requiere que un alumno apruebe el elemento para aprobar el curso) o no ("falso", lo que significa que es una evaluación que le permite al alumno practicar sus habilidades, como un cuestionario de práctica o un elemento que proporciona información, como una clase o lectura, independientemente de que sea un elemento opcional).

2.2. Ingeniería de datos

En este apartado, se llevará a cabo el procesamiento de los archivos seleccionados con el objetivo de obtener el log de eventos que servirá como base para el análisis. A continuación, se detalla el proceso seguido para la obtención y preparación del log de eventos y posterior creación de modelo predicción.

2.2.1. Creación de sesiones

Para la creación de las sesiones de cada usuario dentro del curso MOOC, se obtuvo la información necesaria del dataset “course_progress” donde las principales variables a manipular son: course_item_id, user_id, course_progress_state_type_id, course_progress_ts. Una variable, llamada 'Sesión', fue añadida al conjunto de datos, con valores iniciales configurados en cero, esta variable se actualizó según el cálculo de cada sesión, para el cálculo de dichas sesiones, se apoyó en la obtención del tiempo de inactividad de los usuarios dentro de la plataforma, dicho tiempo, no podía sobrepasar los 40 minutos, lo cual esta en concordancia con los parametros que presento Kovanović (2015) en su estudio.

El algoritmo consiste en ordenar el dataset primero por el campo de estudiante (user_id) seguido por el campo de la marca de tiempo (course_progress_ts). Una vez ordenado el dataset se recorre registro por registro obteniendo la diferencia de tiempo entre actividades (tiempo de la activada k+1 menos el tiempo de la actividad k), si la diferencia de tiempo es mayor de 40 minutos el número de sesión aumenta y es agregada la sesión al campo “Sesion” en el registro k. Es importante mencionar que, si cambia el usuario, se reinicia la sesión a 1; de esta manera se crean las sesiones para cada usuario según el tiempo de inactividad, en la Figura 4 se puede observar un fragmento del dataset con las sesiones creadas.

Figura 4. *Fragmento de los datos creación de sesiones de usuarios. Fuente. Elaboración propia*

	course_item_id	user_id	course_progress_state_type_id	course_progress_ts	Sesion
0	VX8ON	000379b6a2d7d31426ecec469d18023124a051d5	1	2015-11-18 10:10:19.506	1
1	VX8ON	000379b6a2d7d31426ecec469d18023124a051d5	2	2015-11-18 10:12:26.525	1
2	c4BNK	000379b6a2d7d31426ecec469d18023124a051d5	2	2015-11-18 10:12:45.962	1
3	b3oKc	000379b6a2d7d31426ecec469d18023124a051d5	2	2015-11-18 10:12:49.570	1
4	uTSaQ	000379b6a2d7d31426ecec469d18023124a051d5	1	2015-11-18 10:12:53.704	1

2.2.2. Creación de actividades

La creación de actividades consiste en identificar las distintas actividades realizadas y clasificar su tipo, el dataset creado servirá para posterior creación de los estados de las actividades y el análisis de los patrones de aprendizaje, esto se detalla más adelante.

Para la creación de las actividades y su tipo dentro del curso MOOC, se hace uso de dos dataset course_items y course_item_types los cuales se unen mediante el uso de la función "merge" el cual nos da como resultado un dataset "course_item_info" que contiene la información relevante sobre las actividades realizadas por los estudiantes en el curso, En la Figura 5 se observa la una parte del dataset.

Figura 5. *Fragmento de los datos actividades dentro un curso MOOC. Fuente. Elaboración propia*

	course_id	course_item_id	course_lesson_id	course_item_order	course_item_type_id	course_item_name	course_item_optional
0	9uthHBq9EeWg_RJGAuFGjw	ufvxY	K8TJE	0	1	Motor DC	f
1	9uthHBq9EeWg_RJGAuFGjw	1gJgZ	8Pgny	7	3	Ejemplo char y string	t
2	9uthHBq9EeWg_RJGAuFGjw	6AJoc	caqrg	2	1	Prototipo mecánico	f
3	9uthHBq9EeWg_RJGAuFGjw	M1SSE	dkvQP	3	1	Método de mallas	f
4	9uthHBq9EeWg_RJGAuFGjw	VTOmM	mcgiL	9	3	Ejemplo entradas y salidas digitales	t

2.2.3. Creación de estados de las actividades realizadas por los estudiantes dentro del curso MOOC

Para la creación de los estados de las distintas actividades dentro del curso MOOC se crearon dos scripts, de forma inicial se hace uso del dataset creado "course_item_info" y el dataset con las sesiones "course_progress_sesions", se une los dos dataset mediante por el

campo “course_item_id”, una vez creado el dataset este es exportado en formato .csv con el nombre “sesions_items”, Figura 6.

Figura 6. *Fragmento de los datos Estructura del sesions_items. Fuente. Elaboración propia*

	course_item_id	user_id	course_progress_state_type_id	course_progress_ts	Sesion	course_item_type_category	course_item_name
0	VX8ON	000379b6a2d7d31426ecec469d18023124a051d5	1	2015-11-18 10:10:19.506	1	lecture	Video de Bienvenida
1	OE4kr	000379b6a2d7d31426ecec469d18023124a051d5	2	2015-11-19 11:55:52.572	3	supplement	Ejemplo botones arreglo
2	86CMz	000379b6a2d7d31426ecec469d18023124a051d5	1	2015-11-19 11:55:46.817	3	lecture	Arreglos
3	qAlfS	000379b6a2d7d31426ecec469d18023124a051d5	2	2015-11-19 11:55:40.187	3	supplement	Ejemplo botonesIncDec
4	fi6Ba	000379b6a2d7d31426ecec469d18023124a051d5	2	2015-11-19 11:55:31.598	3	supplement	Ejemplo recorre tabla

Después de la creación de los estados de las actividades, se observa que los estudiantes, al no concluir las tareas en un primer intento, las retomaban posteriormente. Asimismo, aunque completaran las actividades, con frecuencia las repetían, y que los estados variaban entre 1 o 2, por lo tanto, al dataset “sesions_items” se agregaron las variables “state_each_item” y “state_each_item_type”. Para “state_each_item” se agregan los estados en base a la columna “course_progress_state_type_id” la cual consiste en los datos 1(inicia) y 2 (completa) y la columna “course_item_type_category” con los tipos de actividades (lecture, supplement y quiz); los estados fueron analizados de la siguiente manera:

- Si el estado era 1 por primera vez → Estado = Inicia.
- Si el estado era 1 por segunda vez → Estado = Vuelve a iniciar.
- Si el estado era 2 por primera vez → Estado = Completa.
- Si el estado era 2 por segunda vez → Estado = Vuelve a repetir.

Al estado se le concatena el tipo de actividad y ese nuevo dato se ingresado en el registro k de la columna “state_each_item”. Se indica una Tabla 1 donde presentan todos los estados posibles según el tipo de actividad.

Tabla 1. *Estados de las actividades lecture, supplement y quiz. Fuente. Elaboración propia*

Estado	Descripción
Inicia Lecture	Estado cuando un estudiante abre una lectura
Vuelva a iniciar Lecture	Lectura que se inicia pero no se completa
Completa Lecture	Una lectura se terminó completamente
Vuelve a repetir Lecture	Una lectura que ya fue completada vuelve a ser abierta
Inicia supplement	Estado cuando un estudiante abre un supplement
Vuelva a iniciar supplement	Supplement que se inicia pero no se completa
Completa supplement	Un supplement se terminó completamente
Vuelve a repetir supplement	Un supplement que ya fue completada vuelve a ser abierta
Inicia quiz	Estado cuando un estudiante abre un quiz
Vuelva a iniciar quiz	Quiz que se inicia pero no se completa
Completa quiz	Un quiz que se terminó completamente
Vuelve a repetir quiz	Un quiz que ya fue completada pero vuelve a ser abierta

De igual manera para la columna “state_each_item_type” se aplicó la misma lógica qué a la primera variable, con la diferencia de qué esta no era concatenada con el tipo de actividad obteniendo como resultado una tabla de estructura similar a la Figura 7.

Figura 7. *Fragmento de los datos Estructura de sesions_items_Estados con los estados por actividad. Fuente. Elaboración propia*

user_id	course_progress_state_type_id	course_progress_ts	Sesion	course_item_type_category	course_item_name	state_each_item	state_each_item_type
1426ecec469d18023124a051d5	1	2015-11-18 10:10:19.506	1	lecture	Video de Bienvenida	Inicia Lecture	Inicia
1426ecec469d18023124a051d5	2	2015-11-18 10:12:26.525	1	lecture	Video de Bienvenida	Completa Lecture	Completa
1426ecec469d18023124a051d5	2	2015-11-18 10:12:45.962	1	supplement	Cuestionario Inicial	Completa supplement	Completa
1426ecec469d18023124a051d5	2	2015-11-18 10:12:49.570	1	supplement	Lista de materiales	Completa supplement	Completa
1426ecec469d18023124a051d5	1	2015-11-18 10:12:53.704	1	lecture	Motivación	Inicia Lecture	Inicia

2.2.4. Creación del log de eventos

La creación del log de eventos se realizará a partir de la unión de los datasets sessions_items_estados y gradebook, utilizando como clave user_id en el primer dataset y

Anonymized Coursera ID (user_id) en el segundo. En esta integración, se conservarán todas las variables del dataset sessions_items_estados, mientras que de gradebook solo se extraerá la variable *Course Passed*, que representa la aprobación o reprobación del curso. La Figura 8 contiene el identificador único de usuario (user_id), la variable objetivo Course Passed, y los registros sobre el progreso de los estudiantes en el curso, como el tipo de actividad (course_item_type_category), el nombre de la actividad (course_item_name), el estado del progreso (state_each_item) y la marca de tiempo asociada a la actividad realizada (course_progress_ts).

Figura 8. *Fragmento de los datos Log de Eventos. Fuente. Elaboración propia*

user_id	Course Passed	course_progress_state_type_id	course_progress_ts	Sesion	course_item_type_category	course_item_name	state_each_item	state_each_item_type
469d18023124a051d5	0.0	1	2015-11-18 10:10:19.506	1	lecture	Video de Bienvenida	Inicia Lecture	Inicia
469d18023124a051d5	0.0	2	2015-11-18 10:12:26.525	1	lecture	Video de Bienvenida	Completa Lecture	Completa
469d18023124a051d5	0.0	2	2015-11-18 10:12:45.962	1	supplement	Cuestionario Inicial	Completa supplement	Completa
469d18023124a051d5	0.0	2	2015-11-18 10:12:49.570	1	supplement	Lista de materiales	Completa supplement	Completa
469d18023124a051d5	0.0	1	2015-11-18 10:12:53.704	1	lecture	Motivación	Inicia Lecture	Inicia

2.2.5. Obtención del 50% del curso

El dataset inicial recopila información sobre las actividades realizadas por los estudiantes a lo largo de todo el curso, dado que el objetivo es desarrollar un modelo predictivo capaz de determinar si un alumno aprobará o no en función de su interacción con la plataforma, se decidió utilizar únicamente el 50% del curso completado, es decir, las primeras dos semanas. Para ello, se identificaron específicamente las actividades realizadas durante este período y se filtró el log de eventos para considerar únicamente los registros correspondientes a las dos primeras semanas. Esta decisión responde a la necesidad de que el modelo pueda hacer predicciones mientras el curso aún está en progreso. Si se entrenara con datos de todo el curso, la predicción solo sería posible una vez finalizado, lo que perdería su utilidad para reconocer tempranamente a los alumnos que podrían reprobar.

2.2.6. Creación de número de actividades realizadas por sesión

Para la creación del número de actividades según el tipo de actividad se hace uso del dataset generado en la creación de sesiones con el 50% del curso llamado df_filtrado50, se

ordena por usuario y sesión, luego se procede a agrupar por usuario, sesion y categoría; en la **¡Error! No se encuentra el origen de la referencia.**Figura 9 se muestra el resultado del agrupamiento.

Figura 9. *Fragmento de los datos agrupados por usuario, sesion y categoría Fuente. Elaboración propia*

	user_id	Sesion	course_item_type_category	count
0	000379b6a2d7d31426ecec469d18023124a051d5	1	lecture	13
1	000379b6a2d7d31426ecec469d18023124a051d5	1	quiz	12
2	000379b6a2d7d31426ecec469d18023124a051d5	1	supplement	2
3	000379b6a2d7d31426ecec469d18023124a051d5	2	quiz	1
4	000379b6a2d7d31426ecec469d18023124a051d5	3	lecture	29

Se pivota la tabla anterior y mediante un contador se crean las actividades realizadas en cada sesión por X usuario, para los valores nulos generados estos son remplazados con ceros, ya que indican qué no realizó dicha actividad en ningún momento. En la Figura 10 se visualiza el resultado final para crear el número de actividades del tipo quiz, lecture y supplement por sesión, obteniendo el nuevo dataset “num_activiades_sesion”.

Figura 10. *Fragmento de los datos num_activiades_sesion Fuente. Elaboración propia*

	user_id	Sesion	lecture	quiz	supplement
0	000379b6a2d7d31426ecec469d18023124a051d5	1	13	12	1
1	000379b6a2d7d31426ecec469d18023124a051d5	2	0	1	0
2	000379b6a2d7d31426ecec469d18023124a051d5	3	28	1	20
3	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	1	3	0	1
4	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	2	0	0	1
...	...	...	...	...	...
45386	ffff6fe3c55bc081e4a4ecf4bcf164c5df38b853	23	5	0	0
45387	ffff6fe3c55bc081e4a4ecf4bcf164c5df38b853	24	3	0	2
45388	ffff6fe3c55bc081e4a4ecf4bcf164c5df38b853	25	5	0	2
45389	ffff6fe3c55bc081e4a4ecf4bcf164c5df38b853	35	0	1	0
45390	ffff6fe3c55bc081e4a4ecf4bcf164c5df38b853	40	1	0	0

2.2.7. Creación de tamaños de sesión por su duración

Para la creación de los tamaños de sesiones por cada estudiante dentro del curso MOOC, se toma en cuenta el tiempo que permaneció un estudiante en cada una de sus sesiones, para ello se eligió el dataset "df_sessions" que contiene el tiempo de duración de un usuario por sesión, consta de todas las sesiones que el usuario tuvo a lo largo del curso y la duración de cada una de ellas en días, horas, minutos y segundos. Dicho dataset consta de 55491 registros y se realizó en base al dataset ya existente "course_progress_sesions", en donde se tomó la diferencia entre el primer tiempo de cada sesión y el último tiempo de la sesión por usuario**.** En base a este dataset se recortó la variable de "duración" y solo se tomó en cuenta los minutos y segundos, en caso de tener horas por sesión estas se las transformó en minutos y se los sumó a los otros minutos, teniendo como resultado la variable "tiempo" que consta únicamente de minutos y segundos por sesión. En la columna "tamano_sesion" existen únicamente 3 valores: "pequena" si el tiempo por sesión en minutos es menor o igual a 10 minutos, "mediana" si el tiempo por sesión en minutos es mayor a 10 minutos y menor a 49 y por último "grande" si el tiempo por sesión en minutos es mayor o igual a 49 minutos. Para clasificar cada sesión por dichos valores se tomó en cuenta la columna "tiempo" y se consideró únicamente solo los minutos, obteniendo así el dataset duración_tamaño_sesion Figura 11.

Figura 11. *Fragmento de los datos duracion_tamano_sesion.*

	user_id	session	tiempo	tamano_sesion
0	000379b6a2d7d31426ecec469d18023124a051d5	1	40.20	mediana
1	000379b6a2d7d31426ecec469d18023124a051d5	2	0.00	pequena
2	000379b6a2d7d31426ecec469d18023124a051d5	3	19.12	mediana
3	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	1	5.17	pequena
4	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	2	14.22	mediana

En la Figura 11 en el caso de la de la sesión del usuario con user_id "000379b6a2d7d31426ecec469d18023124a051d5" en su segunda sesión se marca una duración de 0, ya que después de revisar la fecha y hora indicada 015-11-18 17:17:43.153 el usuario solo registra haber iniciado a una actividad pero nunca la termino(para que cambie de estado) o cambio de actividad para que se inicie otra. Además, se revisó el

archivo num_actividades_sesion.csv y el alumno tenia una sola actividad por lo que se podría suponer que salió de la plataforma.

2.2.8. Creación de duración de sesiones y número de actividades que se haya realizado durante la sesión.

Para saber cuántas actividades (video-lecturas, evaluaciones sumativas y evaluaciones formativas) realizó un estudiante dentro del curso MOOC por sesión, se toma en cuenta dos dataset “duracion_tamano_sesion.csv” y “num_actividades_sesion.csv”, en el primer dataset se consideró únicamente las columnas: “user_id”, “session” y “tiempo”, mientras que en el segundo dataset se consideraron todas las columnas, teniendo así cinco columnas necesarias. Se une los dos dataset mediante por el campo “user_id” y “session” ya que los dos datasets se encuentran ordenados por la columna “user_id”, creamos el dataset “totalActPorSesion” Figura 12 que consta de 45391 registros.

Figura 12. *Fragmento de los datos totalActPorSesion.*

	user_id	session	lecture	quiz	supplement	total_actividades	tiempo
0	000379b6a2d7d31426ecec469d18023124a051d5	1	13	12	1	26	40.20
1	000379b6a2d7d31426ecec469d18023124a051d5	2	0	1	0	1	0.00
2	000379b6a2d7d31426ecec469d18023124a051d5	3	28	1	20	49	19.12
3	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	1	3	0	1	4	5.17
4	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	2	0	0	1	1	14.22
5	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	3	0	0	1	1	0.00
6	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	4	9	0	0	9	38.33
7	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	5	7	10	0	17	60.18
8	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	6	27	8	0	35	122.17
9	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	7	9	0	4	13	80.15
10	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	8	2	0	1	3	33.24
11	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	9	4	0	1	5	21.51
12	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	10	17	14	4	35	115.29
13	0008b09c6ff5d739d22f6cd635c39cfa4cc2d72a	11	18	3	10	31	186.50
14	0009a0968949884cf9c898e02f9e81a01fa6053c	1	19	14	1	34	92.47

2.2.9. Creación de estado de Actividades y Promedio de Tiempo Total

Para la creación este dataset se basó en el dataset “totalActPorSesion” que anteriormente se creó y un nuevo dataset que realizó la fusión del estado de aprobación del curso con el total de las actividades realizadas por los estudiantes.se excluyeron las sesiones con un tiempo de 0 minutos debido a que estas actividades detectadas como ingresos a la plataforma en los

que el usuario no han realizado ninguna activad afectan el cálculo del tiempo promedio ya que no cambiaron de estado en la actividad. En la Figura 13 se tiene unas vista preliminar del dataset resultante “estadoActividadesPromedioTiempoTotal” que contiene 7554 registros donde se encuentra el user_id que será eliminado por que para los posteriores análisis no aportará información, el Course Passed, el total_lecture, total_quiz, total_supplement, total_activities que representan las actividades totales que realizó el alumno en el trascurso del curso, total_sessions y tiempo_promedio que hacen referencia al número total de sesiones al tiempo promedio dedicado a cada sesión trabajo.

Figura 13. *Fragmento de los datos estadoActividadesPromedioTiempoTotal sin user_id*

Fuente. Elaboración propia

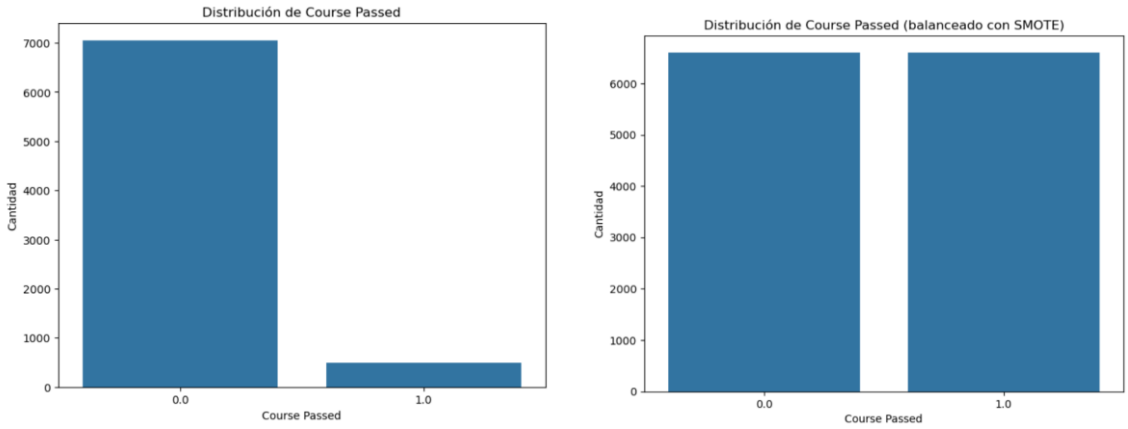
	Course Passed	total_lecture	total_quiz	total_supplement	total_activities	total_sessions	tiempo_promedio
0	0.0	41	13	21	75	2	29.66
1	1.0	96	35	22	153	10	67.68
2	0.0	44	38	9	91	7	31.15
3	0.0	28	0	21	49	1	347.54
4	0.0	71	45	16	132	9	71.71

Con este conjunto de datos se generó una matriz de correlación para identificar relaciones entre variables y se encontró alta correlación entre las variables independientes, lo que sugiere problemas de multicolinealidad. Para mitigar la multicolinealidad, se aplicó Análisis de Componentes Principales (PCA), reduciendo las dimensiones a PCA1 y PCA2, conservando la mayor cantidad de información posible.

Se verificó la distribución de la variable objetivo (CoursePassed) Figura 14(a) y se encontró que se encontraba desbalanceada en relación a la clase 1 que corresponde a los alumnos que aprobaban el curso, para solventar este desbalance se aplicó SMOTE (Synthetic Minority Over-sampling Technique) tras la eliminación de outliers, en la Figura 14(b) se visualiza la nueva distribución de la variable Course Passed después de aplicar SMOTE.

Figura 14. *Distribución de Course Passed . Fuente. Elaboración propia*

(a) (b)



2.3. Ingeniería del modelo de machine learning

Previo al desarrollo del modelo, se realizó un análisis exploratorio utilizando técnicas de aprendizaje no supervisado para identificar posibles clústeres mediante k-means y determinar si existían patrones ocultos en el rendimiento estudiantil. El enfoque de clustering (agrupamiento) se utilizó para explorar la posibilidad de que los estudiantes pudieran agruparse en categorías adicionales, más allá de la simple distinción entre aprobados y reprobados. Para ello, se realizó un análisis del valor óptimo de k, el número de clústeres, utilizando dos métricas: la inercia y el Silhouette Score. La inercia mide la compactación de los clústeres formados, es decir, qué tan cerca están los puntos dentro de cada clúster, mientras que el Silhouette Score proporciona una medida de la calidad del clustering, comparando la cohesión de los clústeres con la separación entre ellos. La Figura 15 muestra cómo la inercia disminuye a medida que aumenta el número de clústeres, donde un valor de k que se estabiliza es un indicio de que se ha encontrado el número adecuado de clústeres, mientras que la Figura 16 indica qué tan bien se separan los clústeres entre sí, con valores cercanos a +1 indicando una buena separación.

Figura 15. Gráfico de Codo mostrando la relación entre el número de clústeres (k) y la inercia. Fuente. Elaboración propia

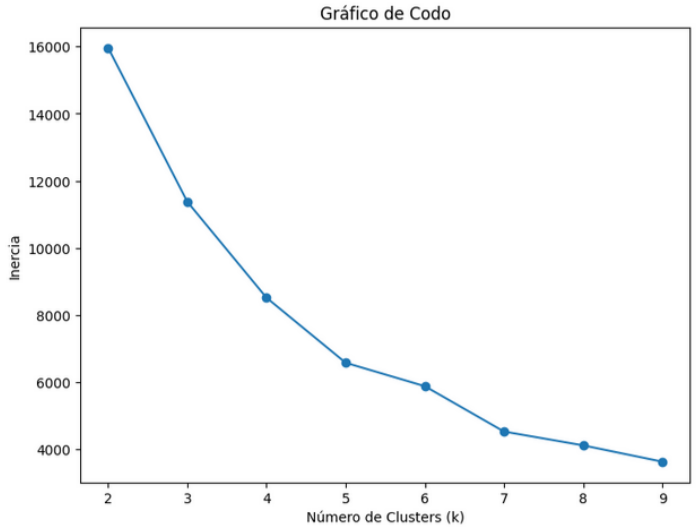
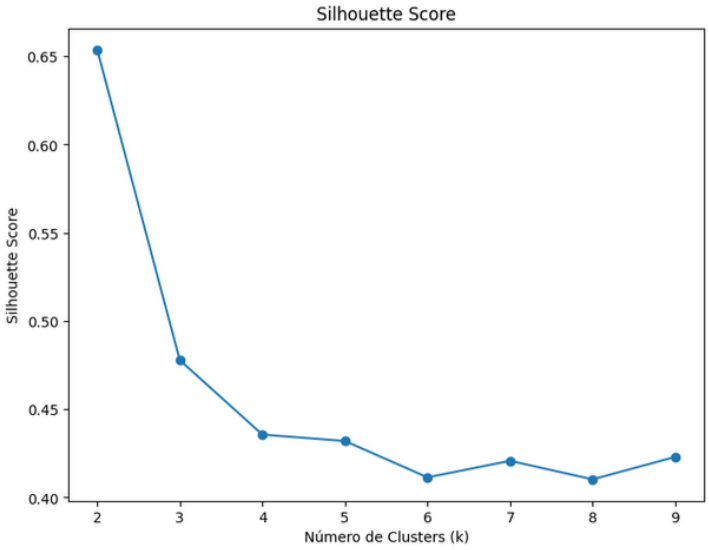


Figura 16. Gráfico del Silhouette Score para evaluar la calidad del agrupamiento en función del número de clústeres. Fuente. Elaboración propia



Tras examinar ambas figuras, se identificó que el valor óptimo de k era 2, lo que corresponde a la existencia de dos clústeres: un grupo de estudiantes con calificaciones

aprobatorias y otro con calificaciones reprobatorias. El análisis de clústeres reveló que, efectivamente, los estudiantes se agrupaban principalmente en dos categorías. No se encontraron categorías adicionales que pudieran diferenciar a los estudiantes dentro de estas dos clases principales, ya que, por las características del conjunto de datos, se identificó que los estudiantes que no registraban estado de aprobación del curso se agruparon claramente en el clúster correspondiente a reprobados. Por lo tanto, el proceso de clustering no aportó descubrimientos nuevos que pudieran enriquecer la clasificación original.

Posterior al análisis realizado se realizó la partición del conjunto de datos en un 80% para entrenamiento y un 20% para prueba, esta partición del conjunto de datos sigue el principio de validación cruzada, que garantiza que el modelo se entrene adecuadamente y se evalúe con datos no vistos durante el proceso de entrenamiento, lo cual asegura una estimación más precisa de su capacidad de generalización y la efectividad de las predicciones para futuros datos. Después se procedió a entrenar y comparar tres modelos de clasificación distintos: Regresión Logística, Árboles de Decisión y Máquinas de Soporte Vectorial (SVM), la selección de estos tres modelos responde a la necesidad de explorar diferentes enfoques y determinar cuál se adapta mejor a las características de los datos.

La Regresión Logística fue elegida debido a su simplicidad y eficacia en tareas de clasificación binaria, como es el caso en este proyecto, donde los estudiantes se clasifican en dos categorías: aquellos que aprueban y no aprueban el curso, al ser un modelo lineal nos permite interpretar con facilidad las relaciones entre las características de interacción de los estudiantes y su rendimiento, lo cual resulta fundamental para identificar qué patrones de participación influyen más en el éxito académico. Los Árboles de Decisión fueron seleccionados por su capacidad para manejar datos con relaciones no lineales y por su habilidad para crear reglas interpretables que explican cómo se toman las decisiones dentro de los cursos MOOC. Dado que las interacciones entre los estudiantes varían considerablemente, como se revisa en la sección 3. 1 sobre Análisis de patrones de interacción, los Árboles de Decisión permiten identificar puntos de corte en el comportamiento estudiantil, los cuales pueden ser indicadores claros de aprobación o reprobación. Mientras que las Máquinas de Soporte Vectorial (SVM) fueron elegidas por su

potencia para manejar grandes volúmenes de datos y su capacidad para clasificar eficazmente en espacios de alta dimensión, siendo útiles cuando los datos no son lineales y permiten encontrar un margen de separación óptimo entre las clases, lo que es crucial para lograr una alta precisión en la predicción del rendimiento académico.

2.4. Aseguramiento de la calidad de las aplicaciones de ML

Para garantizar la efectividad del modelo se analizó la precisión, el recall y el F1-score de cada modelo, además de la visualización de las matrices de confusión para identificar errores en la clasificación. Se seleccionó el modelo con mejor desempeño según las métricas, se verificó la estabilidad del modelo con información no utilizada en el entrenamiento para verificar su capacidad de generalización se exporto a un formato reutilizable para su despliegue en un futuro, asegurando que ofreciera predicciones confiables y precisas.

2.5. Despliegue

El modelo predictivo se implementó en una aplicación que permite a los usuarios ingresar manualmente los datos de las actividades realizadas por los estudiantes para obtener una predicción de aprobación o reprobación del curso, para esto se diseñaron dos interfaces de usuario para facilitar este proceso:

- Interfaz basada en totales donde el usuario ingresa los valores totales de todas las actividades completadas, el número total de sesiones y el tiempo promedio dedicado al curso (Anexo 1).
- Interfaz basada en interacciones específicas aquí el usuario introduce las interacciones en cada actividad realizada, el número total de sesiones y el tiempo promedio dedicado (Anexo 2).

Ambas interfaces aplican los mismos procesos de transformación de datos, incluyendo la normalización de valores y la reducción de dimensionalidad con el modelo PCA entrenado. Posteriormente, el modelo predictivo genera un resultado indicando si el estudiante aprobará o no el curso.

2.6. Monitoreo y mantenimiento del modelo

Una vez que se ha implementado el modelo de predicción para determinar la aprobación o no de los estudiantes en el curso "Electrones en acción", es fundamental establecer estrategias de monitoreo y mantenimiento que permitan garantizar su rendimiento y confiabilidad a lo largo del tiempo. Entre las estrategias de monitoreo y mantenimiento se puede considerar:

- Evaluación continua del desempeño del modelo: la cual consistiría en relazar un seguimiento periódico de las métricas de desempeño del modelo, tales como precisión, recall, F1-score y área bajo la curva ROC en este caso se debería dar la validación con los datos de estudiantes inscritos en ediciones futuras del curso.
- Incorporación de nuevos datos: Periódicamente se debería integrar registros de nuevas cohortes de estudiantes para actualizar el conjunto de entrenamiento.
- Reentrenamiento periódico: si se detecta una desviación significativa en las predicciones, se procederá a reentrenar el modelo con los datos actualizados.
- Ajuste de hiperparámetros: se debería considerar el uso de técnicas como Grid Search o Bayesian Optimization para mejorar la precisión del modelo sin comprometer la generalización.
- Documentación: Se debe mantener una documentación adecuada que permita el proceso de monitoreo y mantenimiento del modelo así si replicabilidad y mejoras futuras.

3. Análisis de resultados

3.1. Análisis de patrones de interacción

A continuación, se presentan los resultados del análisis de patrones interacción de los estudiantes dentro del curso MOOC “Electrones en acción”, para lo cual se responde a las preguntas de investigación planteadas:

P.I.1. ¿Cuáles son las interacciones más frecuentes de los estudiantes en una sesión de estudio en un MOOC?

La pregunta de investigación P.I.1, utilizó la tabla llamada df_patrones, con la cual se realiza un análisis a nivel macro o amplio (hace referencia al conjunto de acciones por

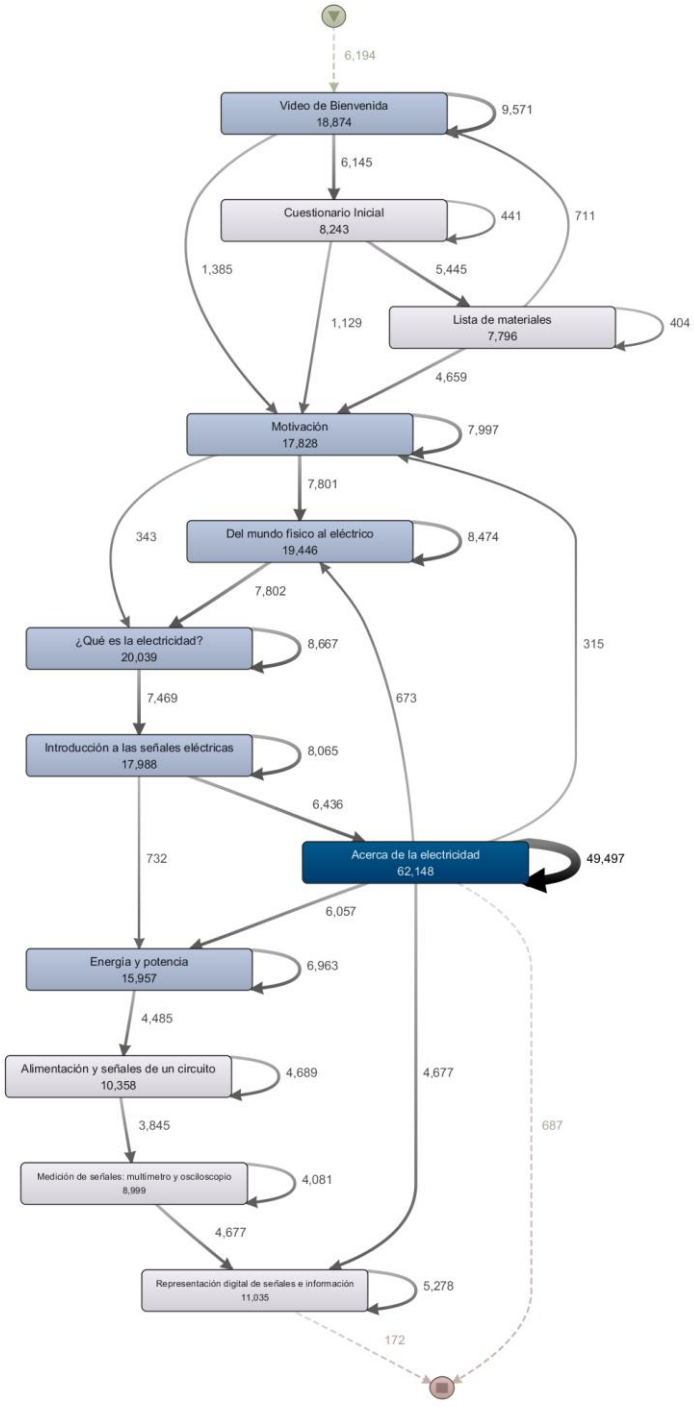
tópico o acciones agrupadas por tema) y a nivel micro o específico (eventos definidos como interacciones con los recursos del MOOC, incluyendo lecturas, suplementarias y evaluaciones, en sus diferentes estados) y se obtuvo que las interacciones más frecuentes de los estudiantes en una sesión de estudio.

Análisis a nivel macro

Se identificaron patrones generales de navegación en el curso, la Figura 17 muestra la secuencia representativa del modelo de procesos, donde cada nodo representa una actividad dentro del curso, y las flechas indican las transiciones entre actividades, aquí el proceso comienza con 6194 sesiones iniciadas, de las cuales 18874 usuarios acceden al "Video de Bienvenida". A partir de esta actividad, se pueden observar diferentes trayectorias: 9571 repiten el video, 6145 avanzan al "Cuestionario Inicial", y 1385 optan por continuar con la actividad de "Motivación". Esto evidencia dos posibles comportamientos: algunos usuarios deciden repetir una actividad antes de continuar, mientras que otros siguen con el flujo natural del curso, avanzando hacia nuevas actividades.

El diagrama facilita la interpretación del comportamiento de los usuarios dentro del curso, permitiendo identificar patrones de navegación, preferencias y posibles puntos de mejora en la estructura del contenido.

Figura 17. *Modelo de procesos a nivel macro. Fuente. Elaboración propia*



Las actividades más frecuentes a nivel macro se presentan en la **Tabla 1**Tabla 2, donde se observa que las 10 actividades más frecuentes durante todo el curso corresponden a actividades de la Semana 1.

Tabla 2. *Actividades más frecuentes a nivel macro. Fuente. Elaboración propia*

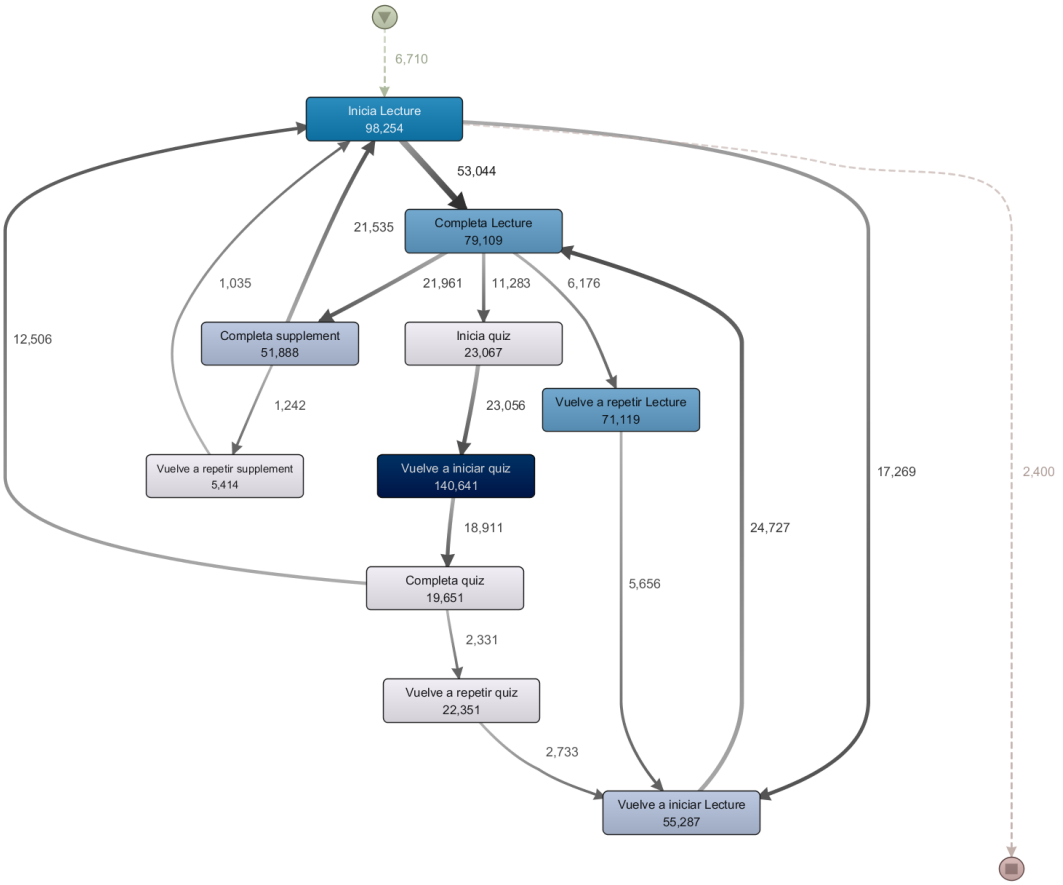
Descripción	**Cantidad de interacciones**
Acerca de la electricidad	62 148
Que es la electricidad?	20 039
Del mundo físico al electrónico	19 444
Video de bienvenida	18 874
Lección 3- modulo 1	18 641
Introducción a las señales eléctricas	17 988
Motivación	17 828
Energía y potencia	15 957
Representación digital de señales e información	11035
Alimentación y señales de un circuito	10358

Análisis a nivel micro

Se analizaron eventos individuales como interacciones especificas con los recursos del MOOC, la Figura 18 muestra una secuencia representativa del modelo de procesos, proporcionando un extracto del modelo general. El proceso inicia con 6710 sesiones, en las cuales la actividad "Inicia Lectura" se realiza un total de 98254 veces. A partir de esta actividad, se pueden identificar tres trayectorias principales: en 53044 ocasiones, los estudiantes completaron la lectura; en 17269 casos, los usuarios decidieron reiniciar la lectura; mientras que, en 2400 ocasiones, los estudiantes abandonaron la sesión tras iniciar la actividad. Después de completar la Lectura los estudiantes se volvieron a dividir en 3 grupos en los cuales 21 961 veces se completó una actividad suplementaria, 11 067 veces se inició una evaluación sumativa y 6176 veces se volvió a repetir una lectura, de los estudiantes que completaron la actividad suplementaria 1242 volvieron a repetir la actividad suplementaria y 1035 inician una nueva lectura, de los que iniciaron una evaluación sumativa 23056 veces vuelven a iniciar la evaluación para posteriormente

18911 veces completar la evaluación, luego de esta actividad toman dos caminos el primero 2331 veces se vuelve a repetir una evaluación y 12506 inician una lectura.

Figura 18. Modelo de procesos a nivel micro. Fuente. Elaboración propia



En este modelo de procesos en la Tabla 3 se muestra que las actividades más frecuentes por los estudiantes a nivel micro son:

Tabla 3. *Actividades más frecuentes a nivel micro Fuente. Elaboración propia*

Descripción	**Cantidad**
Vuelve a iniciar quiz	140 641
Inicia Lectura	98 254
Completa Lectura	79 109

Vuelve a repetir Lectura	71 119
Vuelve a iniciar Lectura	55 287
Completa supplement	51 888
Inicia Quiz	23 067
Vuelve a repetir quiz	22 351
Completa Quiz	19 651
Vuelve a repetir Supplement	5 414

P.I.2. ¿Qué secuencias de aprendizaje se encontraron?

En el curso MOOC Electrones en Acción, se identificaron 6.272 patrones distintos de interacción estudiantil. Un estudio previo, realizado por Mahurad et al. (2018), empleó técnicas de minería difusa para extraer siete estrategias de aprendizaje en cuatro MOOCs de Coursera. Estas estrategias, que se describen a continuación, ofrecen una visión de los comportamientos comunes de los estudiantes en entornos MOOC:

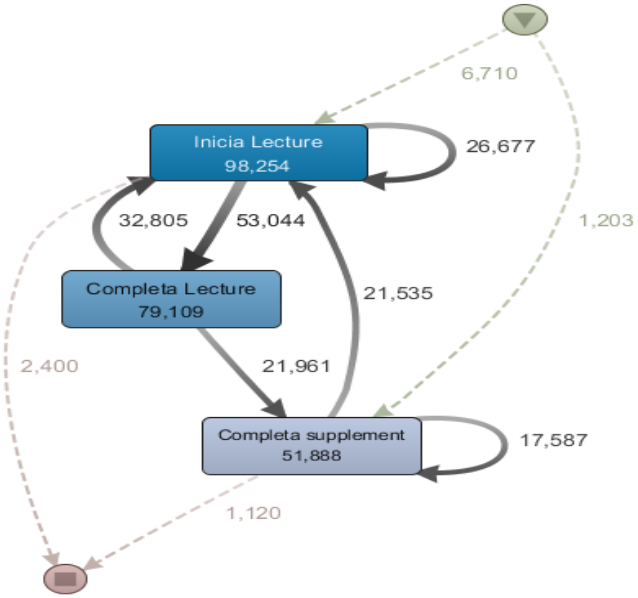
1. Visualización exclusiva de video-lecturas: sesiones donde los estudiantes solo ven video-lecturas, sin interactuar con otros recursos del curso.
2. Enfoque exclusivo en evaluaciones: sesiones en las que los estudiantes se dedican únicamente a realizar evaluaciones, sin consultar materiales de estudio.
3. Visualización completa de video-lectura seguida de intento de evaluación fallido: sesiones en las que los estudiantes ven una video-lectura completa e intentan realizar una evaluación, pero no la aprueban.
4. Intento de evaluación seguido de visualización de video-lectura: sesiones en las que los estudiantes intentan aprobar una evaluación y, al no lograrlo, recurren a las video-lecturas para reforzar su conocimiento.
5. Visualización completa de video-lectura seguida de aprobación exitosa de evaluación: sesiones en las que los estudiantes ven una video-lectura completa y luego aprueban una evaluación.
6. Navegación exploratoria: sesiones en las que los estudiantes exploran el curso sin completar video-lecturas ni evaluaciones, indicando una revisión superficial del contenido.

7. Secuencias de interacción atípicas: patrones de interacción más complejos o inusuales que no se ajustan a las categorías anteriores.

Dado que el curso Electrones en Acción también se encuentra alojado en la plataforma Coursera, es pertinente agrupar los patrones obtenidos dentro de las categorías previamente establecidas en el estudio de Maldonado-Mahauad et al. (2018). Esta clasificación nos permite analizar las estrategias de aprendizaje de los estudiantes dentro de un marco teórico sólido y previamente validado, facilitando comprender como interactúan los estudiantes del curso estudiado.

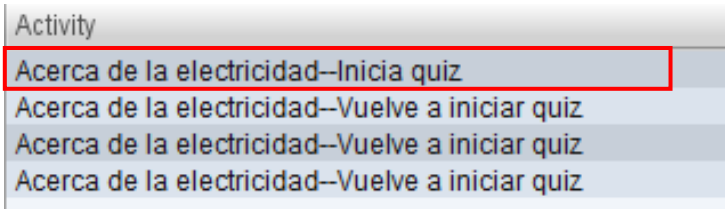
En la Figura 19 se encontraron usuarios inician Lectura, completan lectura, completan una actividad suplementaria y repiten esta secuencia, pero nunca realizan una evaluación.

Figura 19. *Modelo de procesos para patrón lecturas y suplementarias. Fuente. Elaboración propia*



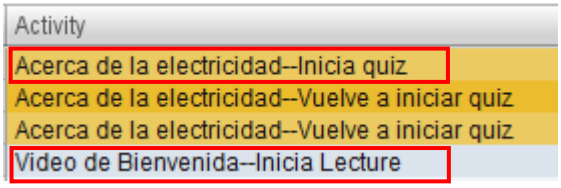
Se encontraron patrones en los cuales los usuarios nunca realizaron una actividad suplementaria o una lectura, estos usuarios se limitaron a realizar solo a realizar evaluaciones sumativas. Aunque existen algunos usuarios los cuales solo inician una evaluación de cualquier módulo, pero nunca la concluyen, y estos después de realizar la actividad abandonan el curso (Figura 20).

Figura 20. *Actividades de usuario que realiza solo evaluaciones. Fuente. Elaboración propia*



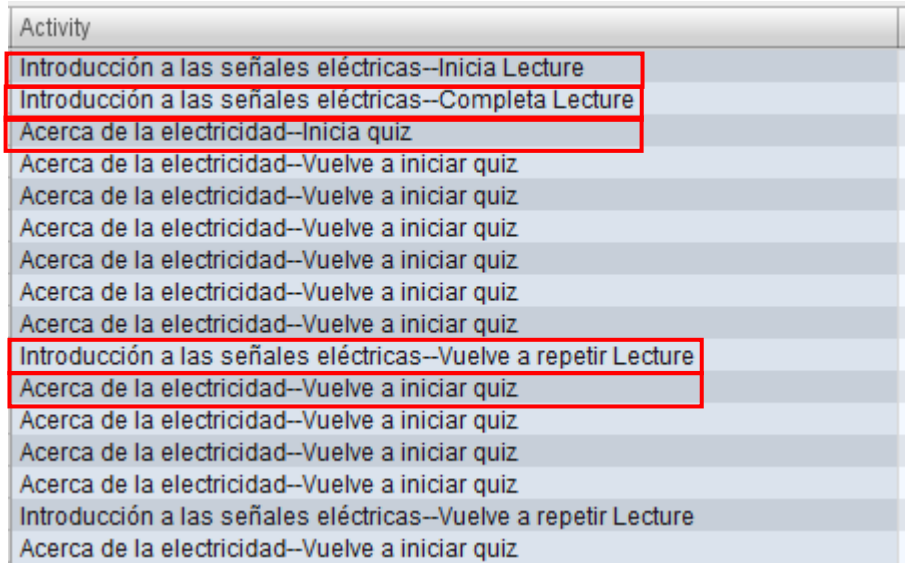
La Figura 21 presenta una secuencia donde los usuarios inician una evaluación (quiz) después al no completar la evaluación inician la lectura que les ayudará a posteriormente a completar la evaluación.

Figura 21. *Actividades de usuario que intenta resolver una evaluación y luego realiza una lectura. Fuente. Elaboración propia*



La Figura 22 presenta una secuencia en la cual se inicia una lectura, se completa lectura, inicia una Evaluación(quiz) después el usuario vuelve a repetir la lectura para así poder contestar la evaluación repite este proceso sin embargo, pero no completa la evaluación.

Figura 22. *Actividades de usuario que completa lectura y luego intenta una evaluación. Fuente. Elaboración propia*



Activity
Introducción a las señales eléctricas--Inicia Lecture
Introducción a las señales eléctricas--Completa Lecture
Acerca de la electricidad--Inicia quiz
Acerca de la electricidad--Vuelve a iniciar quiz
Acerca de la electricidad--Vuelve a iniciar quiz
Acerca de la electricidad--Vuelve a iniciar quiz
Acerca de la electricidad--Vuelve a iniciar quiz
Acerca de la electricidad--Vuelve a iniciar quiz
Acerca de la electricidad--Vuelve a iniciar quiz
Introducción a las señales eléctricas--Vuelve a repetir Lecture
Acerca de la electricidad--Vuelve a iniciar quiz
Acerca de la electricidad--Vuelve a iniciar quiz
Acerca de la electricidad--Vuelve a iniciar quiz
Acerca de la electricidad--Vuelve a iniciar quiz
Introducción a las señales eléctricas--Vuelve a repetir Lecture
Acerca de la electricidad--Vuelve a iniciar quiz

De los 8,017 alumnos de los cuales se tenían registros en el curso MOOC, la herramienta Disco nos permitió identificar 6,710 sesiones de estudio similares, revelando distintos patrones de interacción, se encontraron sesiones en las que los estudiantes solo realizaron video-lecturas sin participar en evaluaciones, así como otras donde se enfocaron exclusivamente en la realización de evaluaciones sin consultar materiales de estudio. También se identificaron casos en los que los alumnos al completar una video-lectura, procedían a realizar una evaluación, pero no lograban aprobarla, o intentaban una evaluación primero y, al no lograr aprobarla, recurrían a la video-lectura para reforzar sus conocimientos. Además, se observaron secuencias en las que los estudiantes completaban una lectura y posteriormente aprobaban la evaluación con éxito, así como comportamientos exploratorios en los que los usuarios navegaban dentro del curso sin completar actividades. Destacan, además, patrones en los que los estudiantes iniciaban y repetían lecturas y actividades suplementarias sin realizar evaluaciones, así como otros en los que solo realizaban evaluaciones sumativas sin interactuar con otros recursos del curso. Finalmente, se identificaron casos en los que los estudiantes intentaban una evaluación, luego consultaban una lectura para reforzar conocimientos, pero no llegaban a completarla, reflejando distintos enfoques en la dinámica de aprendizaje dentro del MOOC. Esta diversidad en las secuencias de aprendizaje refleja la complejidad de los hábitos de los

estudiantes dentro del MOOC, lo que hace que una categorización estricta no capture completamente la variedad de los datos observados.

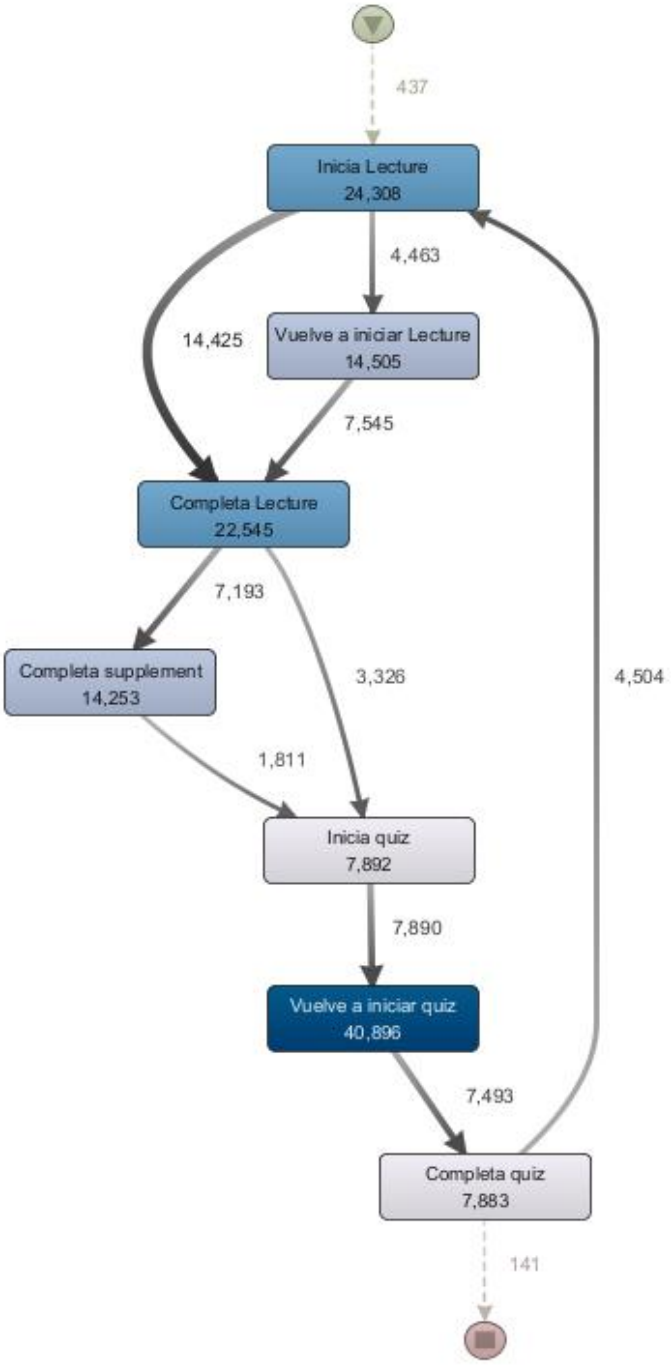
P.I.3. ¿Existen diferencias en las secuencias de aprendizaje entre los estudiantes que lograron aprobar el curso y aquellos que no lo hicieron?

Se observó que los estudiantes que aprobaron el curso tienden a interactuar con el material de manera estructurada como lo presenta la Figura 23, combinando la revisión de lecturas con intentos de evaluación. Sus patrones más comunes incluyen:

- Intentar evaluación → Lectura.
- Iniciar Lectura → Completar evaluación.
- Solo evaluación

Esto sugiere que los estudiantes que tienen éxito en el curso utilizan una estrategia iterativa de aprendizaje, en la que primero intentan resolver una evaluación y luego buscan reforzar su conocimiento a través de los materiales de lectura. Otros, en cambio, revisan los contenidos antes de someterse a una evaluación.

***Figura 23**. Modelo de procesos alumnos que aprueban. Fuente. Elaboración propia*



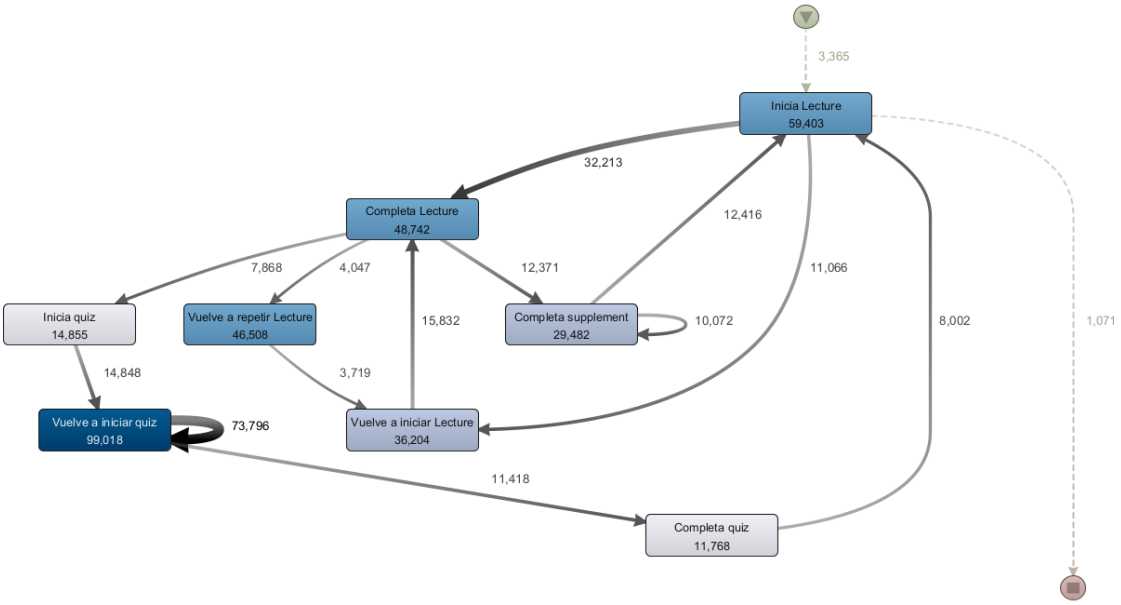
Los alumnos que no aprueban el curso se encuentran dentro del los patrones:

- Solo Lecturas

- Explorar
- Completar Lectura a intentar Evaluación

Estos estudiantes se caracterizan por realizar una menor cantidad de actividades dentro del curso, lo que sugiere una falta de compromiso o dificultades para seguir la estructura del MOOC. Además, en la Figura 24 se identificó que no se seguía una secuencia ordenada de estudio y que algunas actividades se repetían las mismas o volvían a actividades anterior, e inclusive después de una actividad salían dejaban de infatuar, lo que podría haber afectado su rendimiento académico.

Figura 24. *Modelo de procesos alumnos que no aprueban. Fuente. Elaboración propia*



3.2. Evaluación del modelo predictivo

P.I.4. ¿Cómo pueden los modelos predictivos, basados en patrones de interacción, identificar a los estudiantes en riesgo de bajo rendimiento en un curso MOOC?

Para responder a la pregunta de investigación P.I.4, se aplicaron tres metodologías de modelado predictivo, fundamentadas en el análisis de las interacciones de los estudiantes dentro del MOOC. Estos modelos consideraron la frecuencia de participación con los recursos del curso, el tiempo promedio de actividad en la plataforma y el número de sesiones.

Regresión logística

Para determinar la probabilidad de éxito de los estudiantes en el curso (Course Passed), se implementó un modelo predictivo basado en el algoritmo de regresión logística. La tabla que sigue muestra las métricas de evaluación del rendimiento del modelo, derivadas de la aplicación del conjunto de datos de evaluación.

Tabla 4. *Indicadores de evaluación para el modelo de regresión logística. Fuente. Elaboración propia*

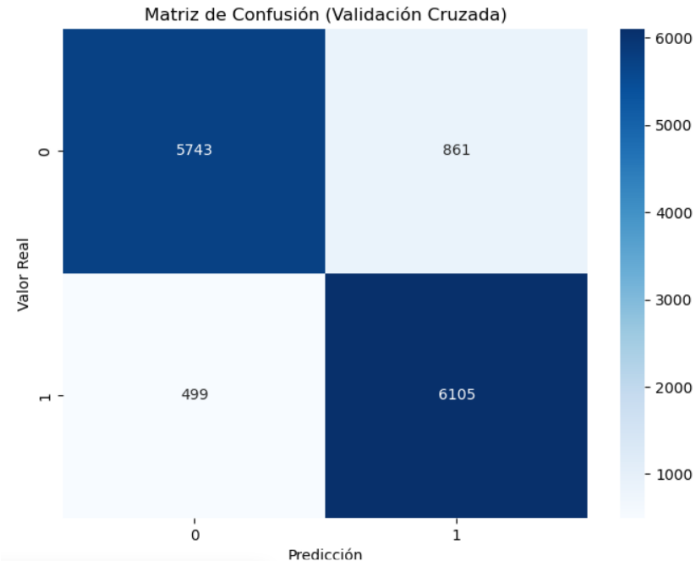
Clase	**Precisión**	**Recall**	**F1-score**	**Soporte**
0.0	0.92	0.87	0.89	6604
1.0	0.88	0.92	0.90	6604
Accuracy	-	-	**0.90**	**13208**
Macro Avg	0.90	0.90	0.90	13208
Weighted Avg	0.90	0.90	0.90	13208
Métricas generales del modelo:				
Accuracy (Precisión global): 0.8970				
Precisión: 0.8764				
Recall: 0.9244				
F1-score: 0.8998				

Las métricas de clasificación para el modelo de Regresion Logistica Tabla 4 muestran que tanto para la clase 0.0 como para la clase 1.0, la precisión, el recall y el f1-score rondan el 90%, lo que indica que el modelo es capaz de clasificar correctamente ambas clases con alta exactitud. La Accuracy (precisión global) del modelo es del 0.897%, lo que significa que el 90% de las predicciones son correctas. Sin embargo, al observar las métricas individuales de precisión, recall y f1-score, encontramos que el modelo tiene una ligera tendencia a clasificar incorrectamente más casos negativos como positivos (falsos positivos) que casos positivos como negativos (falsos negativos). Esto se refleja en una precisión ligeramente inferior (87.6%) en comparación con el recall (92.4%). A pesar de

esto, el modelo mantiene un buen equilibrio entre precisión y recall, como se demuestra en el alto f1-score de 89.9%.

Para analizar la efectividad del modelo, se crea la matriz de confusión, que proporciona información sobre el número de predicciones clasificándolas como correctas o incorrectas.

Figura 25. *Matriz de confusión regresión logística Fuente. Elaboración propia*



En la Para analizar la efectividad del modelo, se crea la matriz de confusión, que proporciona información sobre el número de predicciones clasificándolas como correctas o incorrectas.

Figura 25 se presenta la matriz de confusión revela que el modelo de Regresión Logistica clasifica correctamente una gran cantidad de casos, con 5743 verdaderos negativos y 6105 verdaderos positivos, indicando un buen rendimiento general. Sin embargo, también muestra errores: 861 falsos positivos, donde el modelo predice positivamente cuando es negativo, y 499 falsos negativos, donde predice negativamente cuando es positivo.

Arboles de decisión

Se empleó el algoritmo de árboles de decisión para construir el modelo, con el objetivo de predecir el estado de aprobación de un curso, representado como 'Course Passed'. Los resultados obtenidos, después de aplicar el conjunto de datos de evaluación, se presentan en la siguiente tabla, que muestra las métricas de rendimiento del modelo.

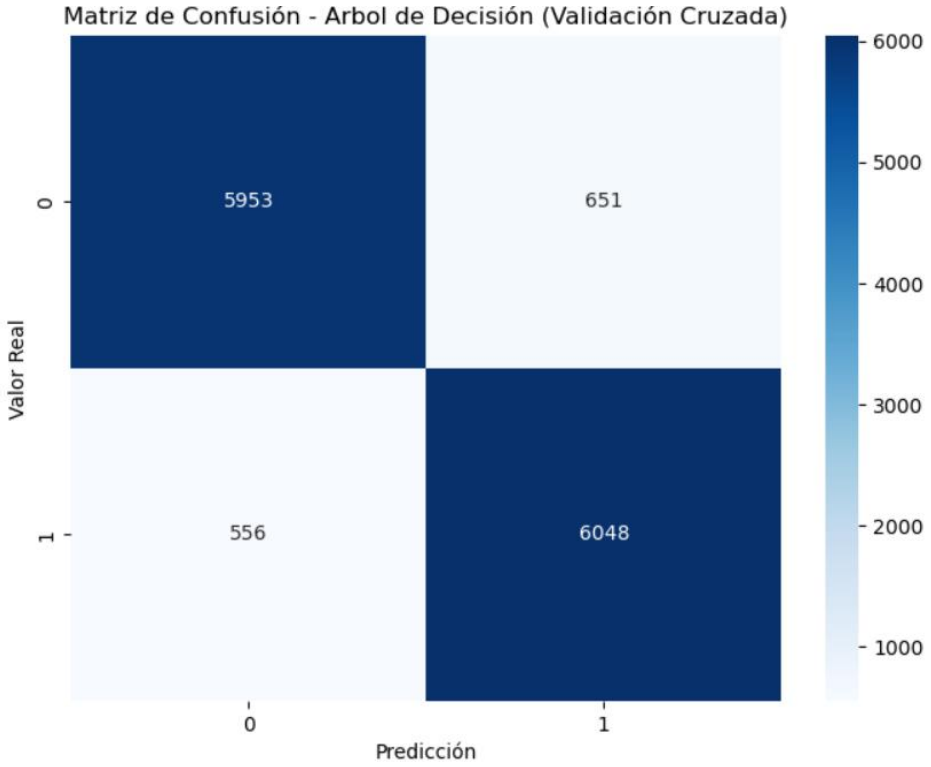
Tabla 5. *Indicadores de evaluación para el modelo de árboles de decisión Fuente. Elaboración propia*

Clase	**Precisión**	**Recall**	**F1-score**	**Soporte**
0.0	0.91	0.90	0.91	6604
1.0	0.90	0.92	0.91	6604
Accuracy	-	-	**0.91**	**13208**
Macro Avg	0.91	0.91	0.91	13208
Weighted Avg	0.91	0.91	0.91	13208
Métricas generales del modelo				
Accuracy (Precisión global): 0.9086				
Precisión: 0.9028				
Recall: 0.9158				
F1-score: 0.9092				

Las métricas de clasificación para el modelo de árbol de decisión Tabla 5 muestran que tanto para la clase 0.0 como para la clase 1.0, precisión, recall y f1-score alrededor del 91%, lo que indica una alta capacidad para clasificar correctamente ambas clases. La Accuracy (precisión global) del modelo es del 91%, lo que significa que el 91% de las predicciones son correctas. El macro promedio y el promedio ponderado también confirman este rendimiento consistente. Aunque hay una ligera diferencia entre la precisión (90.3%) y el recall (91.6%), el modelo mantiene un buen equilibrio entre ambos, como se refleja en el alto f1-score de 90.9%.

Con el propósito de obtener una visión clara del rendimiento del modelo, se genera una matriz de confusión, que muestra la cantidad de predicciones acertadas y las fallidas drl modelo.

Figura 26. Matriz de confusión árbol de decisión



La Figura 26 presenta la matriz de confusión para arboles de decisión revela que el modelo clasifica correctamente una gran cantidad de casos, con 5953 verdaderos negativos y 6048 verdaderos positivos, indicando un buen rendimiento general. Sin embargo, también muestra errores: 651 falsos positivos, donde el modelo predice positivamente cuando es negativo, y 556 falsos negativos, donde predice negativamente cuando es positivo.

Suport vector machine SVM

Se empleó el algoritmo SVM para la creación del modelo, con el objetivo de estimar la probabilidad de que un estudiante apruebe el curso (Course Passed), como resultado, tenemos la siguiente tabla con las métricas de evaluación de rendimiento del modelo, una vez implementado el conjunto de datos de evaluación.

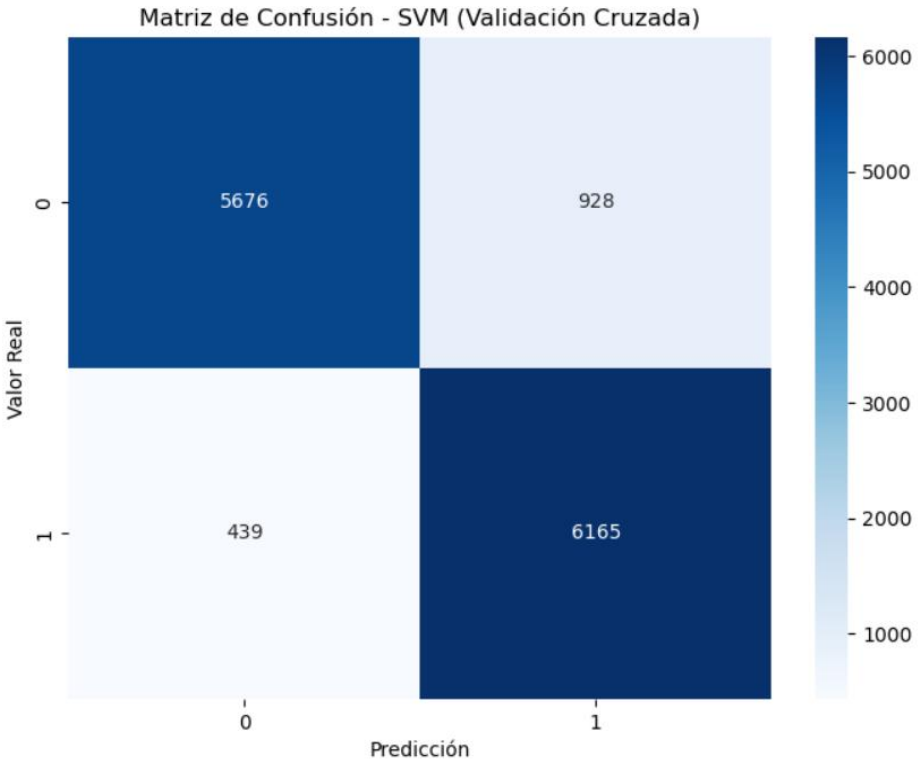
Tabla 6 Indicadores de evaluación para el modelo de SVM. Fuente. Elaboración propia

Clase	**Precisión**	**Recall**	**F1-score**	**Soporte**
0.0	0.93	0.86	0.89	6604
1.0	0.87	0.93	0.90	6604
Accuracy	-	-	**0.90**	**13208**
Macro Avg	0.90	0.90	0.90	13208
Weighted Avg	0.90	0.90	0.90	13208
Métricas generales del modelo				
Accuracy (Precisión global): 0.8965				
Precisión: 0.8692				
Recall: 0.9335				
F1-score: 0.9002				

Las métricas de clasificación para el modelo de SVM Tabla 6 muestran que las métricas de precisión, recall y f1-score alrededor del 90% tanto para la clase 0.0 como para la clase 1.0, lo que indica una alta capacidad para clasificar correctamente ambas clases. La Accuracy (precisión global) del modelo es del 90%, lo que significa que el 90% de las predicciones son correctas. El macro promedio y el promedio ponderado también confirman este rendimiento consistente. Sin embargo, se observa una diferencia notable entre la precisión (86.9%) y el recall (93.4%), lo que sugiere que el modelo tiende a clasificar incorrectamente más casos negativos como positivos (falsos positivos) que casos positivos como negativos (falsos negativos). A pesar de esta diferencia, el modelo mantiene un buen equilibrio entre precisión y recall, como se refleja en el alto f1-score de 90%.

Se crea la matriz de confusión para conocer el número de predicciones de acertadas y fallidas del modelo

Figura 27. Matriz de confusión SVM



La Figura 27 representa la matriz de confusión para SVM que revela que el modelo clasifica correctamente una gran cantidad de casos, con 5676 verdaderos negativos y 6165 verdaderos positivos, lo que indica un buen rendimiento general. Sin embargo, también muestra errores: 928 falsos positivos, donde el modelo predice positivamente cuando es negativo, y 439 falsos negativos, donde predice negativamente cuando es positivo.

En base a los tres modelos revisados el modelo que presenta el mejor desempeño general es el Árbol de Decisión, con un Accuracy de 0.9086 y un F1-score de 0.9093, superando tanto a SVM como a Regresión Logística (Tabla 7). Esto indica que el Árbol de Decisión logra clasificar correctamente a la mayoría de los estudiantes con un buen equilibrio entre precisión y sensibilidad.

Tabla 7. Comparación de métricas de desempeño para los modelos de clasificación

	Regresión Logística	**SVM**	**Árbol de Decisión**
Accuracy	0.897032	0.896502	0.908616
Precision	0.876400	0.869167	0.902821
Recall	0.924440	0.933525	0.915809
F1-score	0.899779	0.900197	0.909269

Si bien el modelo SVM alcanza el valor más alto de Recall (0.9335), su Precision (0.8692) es menor, lo que sugiere que comete más falsos positivos (es decir, clasifica erróneamente a estudiantes como aprobados cuando en realidad no lo están). Además, su Accuracy (0.8965) y F1-score (0.9002) son inferiores a los del Árbol de Decisión, lo que indica que, aunque identifica bien los casos positivos, no logra un buen equilibrio con la precisión.

Por lo tanto, la Regresión Logística muestra un rendimiento similar al de SVM, con una ligera ventaja en Precision (0.8764 vs. 0.8692), pero con un Accuracy (0.8970) y F1-score (0.8998) también por debajo del Árbol de Decisión. Esto sugiere que su capacidad para predecir correctamente los resultados del curso es buena, pero no tan efectiva como la del Árbol de Decisión.

Dando respuesta a la pregunta planteada los estudiantes que aprueban suelen mostrar un mayor número de sesiones, más tiempo promedio en la plataforma y una mayor interacción con videos, lecturas y actividades suplementarias. Por el contrario, aquellos con menor interacción en estos elementos tienden a estar en riesgo de bajo rendimiento, es decir aquellos con baja interacción en recursos clave y menor tiempo promedio en la plataforma suelen ser identificados como de alto riesgo de bajo rendimiento. En este contexto, el modelo de Árbol de Decisión sería la mejor opción, ya que logra un equilibrio óptimo entre Precision, Recall y F1-score, asegurando una clasificación precisa sin comprometer la capacidad de identificar correctamente a los alumnos en riesgo. Además, este modelo tiene la ventaja de ser interpretativo, lo que permite comprender mejor los factores que influyen en el éxito o fracaso de los estudiantes, facilitando la identificación temprana de aquellos que requieren intervención para mejorar su desempeño en el curso.

4. Conclusiones y Recomendaciones

4.1. Conclusiones

Este estudio ha permitido analizar los patrones de interacción de los estudiantes en un MOOC y evaluar la eficacia de modelos predictivos para la identificación de alumnos en riesgo de bajo rendimiento. Los hallazgos más relevantes indican que los estudiantes que aprueban el curso tienden a seguir una secuencia de aprendizaje más estructurada, combinando video-lecturas, actividades suplementarias y evaluaciones, mientras que aquellos que no lo aprueban presentan trayectorias menos organizadas y menor participación en actividades clave.

El análisis de minería de procesos reveló 6,272 patrones de interacción distintos, evidenciando la diversidad de estrategias de aprendizaje. Algunos de estos patrones coinciden con los descritos en estudios previos, lo que valida la relevancia del enfoque empleado. Sin embargo, la gran variabilidad en las secuencias de aprendizaje dificulta una categorización estricta de los estudiantes, resaltando la complejidad de los hábitos de estudio dentro del entorno MOOC.

En cuanto a la predicción del rendimiento académico, se compararon tres modelos de clasificación: Regresión Logística, Máquinas de Soporte Vectorial (SVM) y Árbol de Decisión. Los resultados demostraron que el modelo de Árbol de Decisión alcanzó el mejor desempeño con una precisión del 90.86% y un F1-score de 90.93%, logrando un equilibrio adecuado entre precisión y recall. Este modelo mostró ser una herramienta efectiva para la identificación temprana de estudiantes en riesgo, permitiendo desarrollar estrategias de intervención personalizadas.

Asimismo, se determinó que la frecuencia de interacción con los recursos del curso, el tiempo promedio en la plataforma y el número de sesiones influyen directamente en la probabilidad de aprobación. Los estudiantes con una mayor participación en video-lecturas, actividades suplementarias y evaluaciones tienen más probabilidades de aprobar, mientras que aquellos con menor interacción presentan un riesgo significativo de reprobación.

Si bien los modelos predictivos ofrecen un enfoque prometedor para la detección de estudiantes en riesgo, su implementación en entornos reales plantea desafíos, como la necesidad de acceso a datos en tiempo real y la consideración de factores externos que pueden influir en el rendimiento académico, como la motivación personal y el contexto socioeconómico. La aplicabilidad del modelo en escenarios educativos reales dependerá de la integración efectiva de las herramientas de analítica de aprendizaje dentro de las plataformas de educación en línea.

4.2. Recomendaciones

A partir de los hallazgos de esta investigación, se proponen las siguientes recomendaciones y líneas de trabajo futuro para mejorar la identificación temprana de estudiantes en riesgo y optimizar el diseño y desarrollo de los MOOCs.

Se recomienda la integración del modelo predictivo en la plataforma del curso para monitorear en tiempo real el comportamiento de los estudiantes y generar alertas tempranas que permitan intervenciones oportunas. Para ello, es crucial el desarrollo de estrategias de apoyo personalizadas, como tutorías y recomendaciones de recursos específicos que fomenten la participación activa.

Optimizar el diseño instruccional del curso, asegurando que los materiales sean accesibles, dinámicos y alineados con los objetivos de aprendizaje, además de incentivar la interacción en foros y actividades colaborativas para mejorar la retención.

Monitoreo continuo del modelo predictivo, incorporando nuevos datos y reentrenándolo periódicamente para garantizar su precisión. También se recomienda la exploración de modelos de aprendizaje automático más avanzados, como redes neuronales o modelos de ensamble, que podrían mejorar la robustez del sistema.

La integración de herramientas de analítica de aprendizaje dentro de la plataforma del curso permitirá a los instructores y administradores acceder a información en tiempo real y tomar decisiones informadas. Para ello, es esencial capacitar a los docentes en el uso e interpretación de estas herramientas.

Evaluar el impacto del modelo en un entorno educativo real sin la necesidad de que los estudiantes ingresen manualmente sus interacciones, lo que facilitaría la generación de informes automáticos sobre su desempeño.

5. Referencias Bibliográficas

Aguirre, H. y Rincón, N. (2015). Minería de procesos: desarrollo, aplicaciones y factores críticos. Cuadernos de Administración, 28 (50), 137-157

AWS. (2024). ¿Qué es el análisis predictivo? Recuperado de https://aws.amazon.com/es/what-is/predictive-analytics/

Bernal-González, M. C. (2015). Abandono de los estudiantes en los MOOC.

Bianchi, S., et al. (2022). Cursos online abertos massivos (MOOCs) como potencializadores do conhecimento em instituições de educação superior. Revista Ibérica de Sistemas e Tecnologias de Informação, (48), 59–73. https://doi.org/10.17013/risti.48.59-73

Cáceres, P., et al. (2020). Learning analytics in higher education: A review of impact scientific literature. IJERI: International Journal of Educational Research and Innovation, (13), 32–46.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Ciencia de Datos. (2020). Máquinas de Vector Soporte (Support Vector Machines, SVMs). Recuperado de https://cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines

Coursera. (2019). Recuperado de https://bit4learn.com/es/lms/coursera/

DataCamp. (2024). Tutorial sobre máquinas de vectores de soporte con Scikit-learn. Recuperado de https://www.datacamp.com/es/tutorial/svm-classification-scikit-learn-python

Fincham, O. E., Gašević, D. V., Jovanovic, J. M., & Pardo, A. (2018). From study tactics to learning strategies: An analytical method for extracting interpretable representations. IEEE Transactions on Learning Technologies. https://doi.org/10.1109/TLT.2018.2823317

Fluxicon. (2024). Disco. Recuperado de https://fluxicon.com/disco/

García-Peñalvo, F. J., Fidalgo-Blanco, Á., & Sein-Echaluce, M. L. (2014). Tendencias en los MOOCs.

Geigle, C., & Zhai, C. X. (2017). Modeling MOOC student behavior with two-layer hidden Markov models. In L@S 2017 - Proceedings of the 4th ACM Conference on Learning at Scale. https://doi.org/10.1145/3051457.3053986

Gómez, M. (2022). Analítica del aprendizaje para la visualización y compresión de patrones de estudiantes en edX.

IBM. (2024). ¿Qué es el aprendizaje supervisado? Recuperado de https://www.ibm.com/mx-es/topics/supervised-learning

Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. Internet and Higher Education. https://doi.org/10.1016/j.iheduc.2017.02.001

Kloos, C., et al. (2016). Proyecto eMadrid: MOOCs y Analítica del Aprendizaje. XVIII Simposio Internacional de Informática Educativa, SIIE 2016.

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S., & Hatala, M. (2015). Penetrating the black box of time-on-task estimation. In ACM International Conference Proceeding Series. https://doi.org/10.1145/2723576.2723623

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Muñoz-Gama, J. (2018). Mining theory-based patterns from Big Data: Identifying self-regulated learning strategies in Massive Open Online Courses. Computers in Human Behavior, 80, 179-196. https://doi.org/10.1016/j.chb.2017.11.011

Martín, A. & Ramírez, M. (2016). Los MOOC en la Educación Superior. Un análisis comparativo de plataformas. Revista Educativa Hekademos. https://dialnet.unirioja.es/servlet/articulo?codigo=6280720

ml-ops.org. (2024). CRISP-ML(Q): The ML lifecycle process. Recuperado de https://ml-ops.org/content/crisp-ml

Mukala, M. P., Buijs, J. C. A. M., Leemans, M., & van der Aalst, W. M. P. (2015). Learning analytics on coursera event data, (Simpda).

Ruipérez-Valiente, J. (2020). El proceso de implementación de analíticas de aprendizaje. RIED. Revista Iberoamericana de Educación a Distancia, 23(2), 85-101.

Trigwell, K., & Prosser, M. (1991). Improving the quality of student learning: The influence of learning context and student approaches to learning on learning outcomes. Higher Education. https://doi.org/10.1007/BF00132290

Vega, V. (2016). Minería de procesos de software: una revisión de experiencias de aplicación. Departamento de Ingeniería de Sistemas y Computación, Universidad Católica del Norte, Antofagasta, Chile.

Wu, B. (2021). Influence of MOOC learners' discussion forum social interactions on online reviews of MOOC. Education and Information Technologies. https://doi.org/10.1007/s10639-020-10412-z

Wu, B., & Li, P. (2020). Influence of MOOCs eWOM on the number of registrations and completions. IEEE Access.

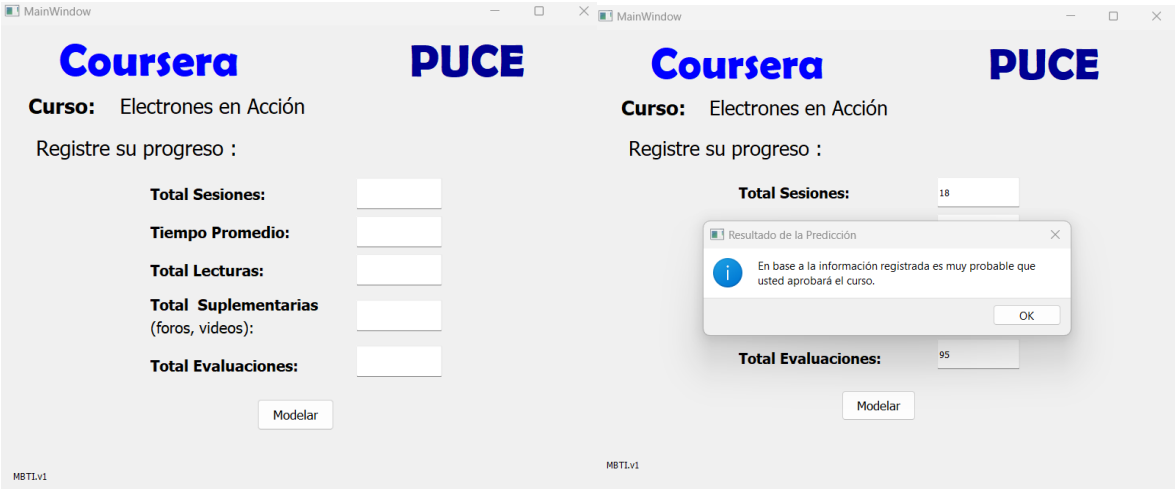
Zapata-Ros, M. (2013). MOOCs, una visión crítica y una alternativa complementaria: La individualización del aprendizaje y de la ayuda pedagógica.

Anexos

Anexo 1

Interfaz basada en totales el usuario ingresa los valores acumulados de todas las actividades completadas, incluyendo el número total de sesiones y el tiempo promedio dedicado al curso. Es importante considerar que para el usuario puede resultar difícil recordar con precisión la cantidad de interacciones realizadas a nivel detallado, ya que no dispone de un desglose específico por tipo de actividad, como lecturas, actividades suplementarias y cuestionarios.

__Figura 28__. Aplicación basada en totales



Anexo 2

Considerando que para el usuario podría resultar difícil o poco realista interactuar de esa manera y obtener una representación más precisa de su participación, se desarrolló una Interfaz basada en interacciones específicas. En esta interfaz, el usuario ingresa, a nivel macro, la cantidad de interacciones que tuvo en cada tipo de actividad realizada, junto con el número total de sesiones y el tiempo promedio dedicado al curso.

Figura 29. Aplicación basada en actividades