Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

Análisis del rendimiento académico en un curso virtual de posgrado de una universidad pública

Aplicación de learning analytics en plataforma Moodle

AUTOR: GABRIEL ALEJANDRO DUARTE

TUTOR: PATRICIA GIRIMONTE

SEPTIEMBRE 2023

Resumen

Este trabajo tiene como objetivo construir indicadores y modelos predictivos del rendimiento académico de los estudiantes, aplicando algunas técnicas de learning analytics, que permitan detectar patrones relacionados con la interacción de los alumnos en el aula virtual de una asignatura de un posgrado.

Para llevar adelante este trabajo, se utiliza un conjunto de datos anonimizados, aportado por las autoridades del posgrado, referido a los registros de la interacción de los estudiantes en una asignatura con las distintas herramientas que ofrece el aula virtual Moodle (Modular Object Oriented Developmental Learning Environment).

Se considera una variable de interés creada a partir del conjunto de datos, la cual, resulta una variable de dos categorías, cada categoría refleja el rendimiento académico del alumno, lo que para este trabajo es si aprueba o no la asignatura del posgrado. Para la obtención de las variables a utilizar en este trabajo, se aplican algunas técnicas de minería de datos utilizando RapidMiner. La base obtenida está formada por variables que reflejan algunos comportamientos de los alumnos en el aula virtual de la plataforma Moodle. Se realizan algunos métodos de clusterización, para encontrar patrones de los estudiantes en foros y otras actividades. Además, se realizan diferentes modelos de Regresión Logística, para clasificar a los alumnos en las dos categorías de la variable rendimiento académico. Se logra encontrar un modelo con una performance del 88,88%, el cual permite identificar variables regresoras, como las que quedan definidas a partir de los foros de consulta y número de conexiones a la plataforma, lo que posibilita concluir que una mayor participación en estos dos recursos lleva a un mejor rendimiento académico. Mediante la aplicación de librerías de R, se logran los resultados de la clusterización y Regresión Logística previamente mencionados.

Palabras clave

Posgrado virtual, Universidad nacional, Rendimiento Académico, Learning analytics, Plataforma Moodle.

Índice

Introducción

Learning analytics o analítica de aprendizaje refiere a la recolección y análisis de datos sobre los estudiantes y sus entornos, con el propósito de comprender y mejorar los resultados del aprendizaje. El punto de encuentro entre la estadística cuantitativa y cualitativa tradicional, utilizadas por las universidades por años, con las herramientas usadas para el procesamiento de grandes volúmenes de datos (Big Data), es precisamente learning analytics.

El uso de learning analytics posibilita convertir la enorme cantidad de datos obtenida desde la plataforma Moodle en información valiosa que permite proponer nuevos modelos de enseñanza y aprendizaje para mejorar el rendimiento académico (Iglesia, 2019).

El interrogante a resolver será: ¿Qué elementos asociados al uso del aula virtual por los estudiantes son indicadores o predictores del rendimiento académico?

El objetivo general relacionado con esta pregunta es la detección, estudio y análisis, mediante el uso de algunas técnicas de learning analytics, de patrones que resulten de la participación de los alumnos en el aula virtual como indicadores o predictores del rendimiento académico de los estudiantes. El logro de este objetivo se llevará a cabo en tres apartados, que se corresponderán con la estructura de este trabajo.

En el primer apartado, se procederá a describir los conceptos de learning analytics, su importancia y alcance. Los datos para este estudio fueron entregados por las autoridades del posgrado, en forma anonimizada, enmarcados en la ley correspondiente para mantener la privacidad de los alumnos. En el conjunto de datos, la mayoría de los mismos, no se presentan de manera bien ordenada y prolija, por este motivo, se necesita una preparación adecuada para su procesamiento. Esta etapa de preparación y adecuación de los registros en la base de datos será una parte importante de este trabajo.

En el segundo apartado, se aplicarán a las variables cuantitativas y cualitativas obtenidas y de interés a partir del conjunto de datos, algunas técnicas de machine learning, como ser, clusterización, minería de textos y otras propias de learning analytics. Se realizarán gráficos y tablas para conseguir una mejor interpretación de los resultados obtenidos. Para llevar a cabo esta tarea se usarán librerías del software libre R, junto a la herramienta de minería de datos RapidMiner. Esto permitirá construir modelos, métricas y comparaciones para utilizar en la etapa siguiente del trabajo.

Por último, en el tercer apartado, se obtendrán y explicarán las variables regresoras que mejor describen el rendimiento académico de los alumnos, lo que permitirá obtener métricas e indicadores para enriquecer no sólo la parte académica por parte del alumno, sino también presentar una oferta superadora en planes de estudios para la gestión del posgrado. Será de vital importancia en este estudio poder encontrar un modelo para predecir si un alumno aprueba o no la asignatura del posgrado, lo que refleja la variable del rendimiento académico.

A partir de todas las etapas anteriores, se obtendrá un posible modelo a implementar, que permita realizar predicciones futuras.

1. Learing analytics y registros obtenidos de un curso de posgrado virtual de una universidad nacional

La incorporación de learning analytics, en el análisis de la educación virtual, adquiere gran importancia para realizar un seguimiento de logros y dificultades de los alumnos de manera individualizada, para así obtener una mejora en la producción académica en materias dictadas en modalidad virtual en asignaturas universitarias. Esta técnica de minería de datos educacional agiliza el procesamiento, comprensión y estudio de los grandes volúmenes de datos que los alumnos generan a partir del uso en distintas actividades de las aulas virtuales de las plataformas. El empleo, análisis y mención de los datos obtenidos por el uso de la plataforma Moodle, requieren protección de privacidad en cumplimiento de la ley 25.326. Este apartado refiere en primer lugar a definir learning analytics. En segundo lugar, se presentan distintas formas de anonimizar datos para ser encuadrados bajo la ley 25.326. Para finalizar el apartado, en tercer lugar, se presenta y explica el conjunto de datos anonimizado entregado por las autoridades del posgrado, y el procesamiento realizado en el mismo para obtener un conjunto de datos de interés para este trabajo.

1.1. Learning analytics y optimización en la enseñanza en posgrados virtuales universitario

Las universidades nacionales de la República Argentina, vistas como organizaciones, intentan siempre aplicar procesos para la mejora de la educación. En los últimos años, tanto en cursos de grado como de posgrado, el uso de las tecnologías ha sido de gran utilidad para intentar conseguir esta mejora y, además, como apoyo de los estudiantes en sus diferentes asignaturas. El uso de learning analytics, junto a la gestión y procesamiento de grandes volúmenes de datos generados por los estudiantes, contribuyen notablemente en la manera en que las universidades nacionales dan seguimiento, e intentan predecir o clasificar el desempeño de sus alumnos.

En la primera Conferencia Internacional de learning analytics, esta analítica de aprendizaje se definió como la medición, recolección, análisis y presentación de datos sobre los estudiantes y sus contextos con el propósito de comprender y optimizar el aprendizaje y los entornos en que

se produce (LAK, 2011). En este sentido, el fin es mejorar los procesos de enseñanza y aprendizaje, para así dar repuestas a las necesidades de los estudiantes.

Algunos autores, entre ellos Van-Barneveld et al. (2012) y García et al. (2018), definen learning analytics como el uso de técnicas para orientar los recursos educativos, curriculares y de apoyo, para el logro de objetivos de aprendizaje específicos. Dentro de las técnicas analíticas se encuentran las estadísticas tradicionales, sistemas de aprendizaje de reglas, minería de texto, aprendizaje computacional, estocásticas, basadas en casos, entre otras, que se aplican en el contexto educativo para analizar el comportamiento de los estudiantes, y poder predecir algún evento futuro desde lo académico hasta lo administrativo o gestión.

Del mismo modo, los autores Dietz-Uhler y Hurn (2013), conciben a learning analytics como la medición, acumulación y reporte de datos relacionados con los estudiantes y su contexto, con el objetivo de entender y optimizar la enseñanza y el ambiente en que se da. Por otro lado, los grandes volúmenes de datos (Big Data) se han descrito como la capacidad de almacenar y manipular grandes cantidades de datos; la información almacenada puede descubrir patrones de desempeño estudiantil y así, sugerir acciones para mejorar el rendimiento académico (Picciano, 2012).

Learning analytics es un área estrechamente ligada a la minería de datos educativa, puesto que posibilita el análisis de grandes volúmenes de datos generados por los estudiantes obtenidos de diversas fuentes, para encontrar patrones ocultos que favorezcan la toma de decisiones basada en información, a fin de tomar acciones que prevengan alguna situación futura tanto en el contexto académico como en el administrativo en una institución de educación superior (Urbina-Nájera, 2021).

En aplicaciones concretas, los autores Arnold y Pistilli (2012), presentaron un estudio para predecir el rendimiento de los estudiantes usando técnicas estadísticas como una introducción de learning analytics. Se basaron en las calificaciones, características demográficas, historial académico y esfuerzo de los estudiantes medidos con la plataforma Blackboard. A partir de este estudio y de los resultados obtenidos, se consiguió un modelo exitoso para disminuir el índice de deserción.

También, Yu y Jo (2014), realizaron un estudio en una universidad de mujeres de Corea del Sur, con una muestra de 84 estudiantes de una licenciatura presencial donde se emplea la plataforma Moodle para descargar material académico. Analizaron 6 atributos, a saber: frecuencia de entrada a la plataforma, tiempo de estudio en la plataforma, regularidad de intervalos de aprendizaje en la plataforma, número de descargas de material, interacción con

compañeros e interacción con el profesor. Mediante la aplicación de regresión lineal múltiple, se obtuvo un modelo para predecir el logro académico de cada estudiante.

Finalmente, Lu et al. (2018), muestran un estudio con datos de 33 hombres y 26 mujeres de un curso virtual de cálculo. Recolectaron 21 variables, entre las que se encuentran: número de videos que el estudiante consulta, número de clicks, número de unidades que el estudiante estudia por semana, entre otras. Aplicando técnicas de minería de datos identificaron que las variables más importantes en ese estudio para determinar el desempeño del alumno fueron: número de días que el estudiante tiene actividad por semana, número de actividades por semana en la cual el estudiante se involucra, número de videos que el estudiante ve completamente, número de clicks que hace por semana y número de veces que el estudiante hace click en pausa.

En este trabajo se aplicarán algunas técnicas de learning analytics a una materia de un posgrado virtual de una universidad nacional, con el objetivo de encontrar indicadores y modelos que permitan predecir el rendimiento académico de los cursantes. Estos indicadores y modelos pueden ser aprovechados y explotados, tanto por el alumno como también por los profesores de la materia y las autoridades del posgrado, para encontrar nuevas estrategias de enseñanza y aprendizaje superadoras a las utilizadas, y así, mejorar el rendimiento académico de los estudiantes.

A continuación se ponen de manifiesto algunos beneficios para los distintos actores involucrados en el dictado de la materia.

Entre los beneficios para el alumno se puede destacar el uso que realiza el estudiante de las herramientas tecnológicas para su aprendizaje, conocer su curva de aprendizaje, su avance, qué aspectos debe mejorar para alcanzar los objetivos planteados por los profesores, ver devoluciones de participaciones en tareas conjuntas e individuales, estas últimas compararlas con las de sus compañeros, cuáles son los recursos más utilizados por sus compañeros, las respuestas obtenidas a sus consultas a través de los foros por parte de los profesores, entre otras. Esta participación del alumno de manera activa para la mejora de su aprendizaje en el contexto de learning analytics se llama aprendizaje autorregulado.

Entre los beneficios para los profesores, podemos remarcar que existe un sinfín de estrategias tecnológicas que pueden mejorar su labor docente. Estas, además, resultan de gran ayuda para entender los diferentes ritmos de aprendizaje de sus alumnos. Mediante la aplicación de técnicas de learning analytics, el profesor tiene a su disposición la posibilidad de observar

patrones en el comportamiento de los alumnos en el aula virtual, lo cual permite un seguimiento personalizado de cada estudiante y favorece, además, a un mayor control del rendimiento académico. Estos patrones ayudan a comprender cuáles son las interacciones que se producen entre los alumnos, y de esta manera poder detectar influencias de algún alumno sobre el resto, comunicación entre los alumnos vía el foro de colaboración entre alumnos, detectar cuáles son los estudiantes que trabajan mejor juntos, o incluso si algún estudiante queda aislado. Este tipo de forma de enseñar se lo conoce como aprendizaje personalizado en el contexto de learning analytics.

Con respecto a los beneficios para la universidad, mediante la implementación de learning analytics, se pueden obtener y analizar los logros de los estudiantes en sus estudios, y esto resulta de mucho valor pues la universidad tiene una constante preocupación por la validación de la calidad de sus procesos de enseñanza y aprendizaje. Una vez que son procesados los datos, interpretados y publicados a las autoridades correspondientes, los mismos son de gran importancia pues proporcionan, por ejemplo, la posibilidad de realizar cambios en planes de estudios, construir nuevas estrategias de enseñanza, implementar nuevos canales de comunicación con el alumno, bajar el porcentaje de deserción, saber el nivel de satisfacción de los estudiantes con sus cursos, evaluar la relación entre alumnos y profesores.

Cabe destacar que la implementación de learning analytics tiene sus desafíos. En primer lugar, un desafío tecnológico, como por ejemplo las herramientas con las que se cuentan y cómo las mismas operan entre sí, además de la naturaleza de los datos que se recolectan. La forma en que se obtienen y almacenan los datos para el uso de técnicas de learning analytics es fundamental. Otros retos están relacionados con el uso de los datos, la ética con la cual son manipulados, con el fin para los cuales son recolectados, la calidad de los datos para usarlos como evidencia de la mejora del aprendizaje, entre otros.

1.2. Ley de datos personales 25.326

La ley de datos personales 25.326 de la República Argentina, sancionada y promulgada parcialmente en el año 2000, tiene como objetivo la protección integral de los datos personales asentados en archivos registros, bancos de datos u otros medios técnicos de tratamiento de datos, sean estos públicos o privados destinados a dar informes, para garantizar el derecho al honor y a la intimidad de las personas, así como el acceso a la información que

sobre las mismas se registre, de conformidad a lo establecido en el artículo 43, tercer párrafo de la Constitución Nacional (art.1º).

La ley es comprensiva además de los denominados datos sensibles, entendidos como aquellos datos personales que revelan origen racial y étnico, opiniones políticas, convicciones religiosas, filosóficas o morales, afiliación sindical e información referente a la salud o a la vida sexual (Travieso & Moreno, 2006).

La información a la que se accedió para la realización de este trabajo tuvo un proceso de anonimización, por parte de las autoridades del posgrado, para proteger la información sensible y de identificación de los casos empleados. El tratamiento y explotación de grandes volúmenes de datos de información pueden ofrecer múltiples beneficios a las universidades, siempre que se encuadren dentro de las leyes que rigen los derechos a las personas, su privacidad y la protección de sus datos personales.

En un proceso de anonimización es fundamental valorar los riesgos de una reidentificación a posteriori y cómo se va a garantizar la confidencialidad de la información personal anonimizada.

Existen diferentes técnicas para la anonimización de datos, éstas se dividen en generales y particulares (Agesic, 2017). A continuación, se exponen siete de las maneras de anonimizar más utilizadas.

La aleatorización es un conjunto de técnicas que modifican la veracidad de los datos con el fin de eliminar el vínculo existente entre ellos y su titular. Es decir, si los datos se vuelven ambiguos, no se puede identificar a una persona concreta.

La adición de ruido es una técnica muy utilizada, la cual resulta de la modificación del conjunto de datos para que sean menos exactos, conservando no obstante su distribución general.

La permutación consiste en mezclar los valores de los atributos en una tabla para que algunos de ellos puedan vincularse artificialmente a distintos interesados. Esta técnica es útil en el caso de que sea importante conservar la distribución exacta de cada atributo en el conjunto de datos.

Las técnicas de agregación y de anonimato k tienen el objetivo de impedir que un interesado sea singularizado cuando se le agrupa junto con, al menos, un número k de personas.

La supresión es un método en el cual algunos valores de los atributos son reemplazados por un NA (Not Available) o eliminados.

En la técnica de generalización, los valores individuales de atributos son reemplazados por una categoría más amplia.

Los registros obtenidos por las autoridades del posgrado a partir de la plataforma Moodle fueron entregados anonimizados para este trabajo, encuadrados bajo la ley 25.326. De esta manera, no se tendrá acceso a nombre y apellido, dirección de IP de las computadoras usadas por los alumnos, fecha de nacimiento, nacionalidad, idioma, lugar de residencia, número de celular y dirección de correo electrónico.

Cada alumno, para su anonimización, tiene asociado un número conocido solamente por las personas involucradas en la obtención de la base de datos, como así también, está codificado el nombre de la asignatura y los nombres de los profesores, para evitar de esta manera la identificación del individuo.

1.3. Descripción de la generación y calidad de los datos. Atributos obtenidos

Los datos para el presente trabajo fueron obtenidos del aula virtual de una asignatura en la plataforma Moodle, teniendo en cuenta los logs de ingreso de los estudiantes a los distintos contenidos teóricos y prácticos. La información obtenida corresponde a los estudiantes de una cohorte de un posgrado virtual. Se obtuvieron alrededor de más de 11000 registros, los cuales fueron almacenados en formato CSV (archivo de valores separados por coma) y anonimizados por parte de las autoridades del posgrado. La recolección de datos es la huella digital que deja cada alumno en el aula virtual en la plataforma Moodle de la materia.

La plataforma provee, en general, la siguiente información para cada interacción de un usuario con la misma: fecha y hora de conexión, número de identificación del usuario, usuario afectado (curso o materia correspondiente), contexto del evento (lo realizado por el usuario), componente (sistema), nombre y descripción del evento (curso visto o foro), origen (número de identificación del usuario y curso visto).

Además de la información conseguida en la plataforma, también en el conjunto de datos está el rendimiento académico de cada alumno, el atributo de estudio en este trabajo, que refleja la aprobación o no de la asignatura.

Los registros relacionados con las actividades de los docentes en la plataforma, como ser carga de material, conexiones para responder consultas en foro, carga de videos, producción de archivos y otras actividades docentes, no estuvieron disponibles, ya que estos registros no

reflejan el comportamiento de un alumno en la plataforma, sólo hacen referencia a actividades propias del docente.

A los atributos de texto, como intercambios en foros o consultas sobre la materia, ya sean de contenido por parte de los profesores o gestión por parte de organización, se les aplicó minería de texto, a través de RapidMiner, para darle un formato compatible con el que se necesita para aplicar técnicas de learning analytics.

Las encuestas realizadas por alumnos de la materia cursada son de gran importancia para saber cuán útil fue para el estudiante el material especialmente preparado por los docentes para los cursos virtuales, además de permitir extraer un estudio del análisis de sentimiento, área de investigación que se encuentra en rápido crecimiento (Florit et al, 2020). La opinión del alumno sobre el docente, si bien es de importancia para las autoridades del posgrado, no lo es para el estudio del comportamiento de un estudiante durante su cursada virtual, como sí lo es la accesibilidad a la información que necesita, la claridad de las respuestas que consiguió a través del foro, como así también, saber si el recorrido por la plataforma para la realización de las distintas actividades fue óptimo. En este trabajo no se contó con las encuestas de opinión de los alumnos.

2. Aplicación de métodos estadísticos, data mining y learning analytics

En este apartado se analizarán y procesarán los datos de la base entregada por las autoridades del posgrado virtual, a la que se hizo referencia en el apartado anterior, con el fin de poder construir indicadores y modelos del rendimiento académico. En esta etapa, además, se generarán nuevas variables a partir de las ya exisitentes en el conjunto de datos disponible. A las variables obtenidas se les aplicarán algunas técnicas de data mining y clusterización.
Para llevar a cabo estas tareas, se aplicarán librerías del software libre R, como así también, la herramienta RapidMiner para el análisis de textos.

2.1. Limpieza y descripción de atributos

Los registros con los que se va a trabajar en esta sección son los proporcionados por las autoridades de un posgrado virtual. El conjunto de datos disponible representa los logs de entrada al aula virtual de una asignatura en la plataforma Moodle durante un cuatrimestre, con una duración de 23 semanas, en el que se dictó la materia. La cantidad de registros, es decir, el número de logs de conexión de los alumnos a la plataforma es 11782 y para cada uno de ellos se dispone de 8 atributos que reflejan la actividad de los estudiantes en el aula virtual.
A continuación se realiza una descripción de las variables con las que se cuenta para este trabajo, los valores de estas variables se obtienen directamente de la plataforma.
Las variables o atributos son; Fecha y hora: es la fecha y hora en la que se efectuó el log de ingreso de la persona; Nombre completo del usuario: este campo está codificado con números por parte de las autoridades del posgrado para evitar la identificación del usuario, en este caso, el alumno; Usuario afectado: número de alumno, código de materia; Contexto del evento: en este atributo está presente lo efectuado por el usuario, y es la descripción del log de entrada, a saber, curso visto, actividad vista, módulos creados, foros de consulta, foros de colaboración, foros de avisos, entregas de actividades adicionales por parte de los alumnos, si hubo una retroalimentación, entre otros eventos; Componente: en esta variable se almacena si el contexto de evento es una tarea o una consulta administrativa; Nombre del evento: representa la sección del curso vista por un usuario; Descripción: en este caso web; Origen: web y número del evento.

El atributo “Nombre completo del usuario” tiene valores del 1 al 18, ya que 18 fue el número de alumnos que participaron en el cursado de la asignatura.

El criterio de éxito para aprobar la materia fue la presentación de dos entregas obligatorias, y para la aprobación de la asignatura se exigió la aprobación de ambas entregas.

Para cada alumno, en este trabajo, se estudió su comportamiento en la plataforma permitiendo generar nuevas variables, basadas en las distintas actividades llevadas a cabo por el mismo.

Para realizar esta tarea, se aplicó sobre la variable original “Contexto del evento” minería de texto mediante el uso de RapidMiner. En este atributo estaba concentrada toda la información de las actividades de cada alumno, por lo que fue de gran importancia detallar y separar las mismas. Por ejemplo, para un log de un día y fecha de un alumno, en la variable “Contexto del evento” aparece si fue una actividad en foro, una entrega de actividad, un módulo de curso visto, entre otras. Para hacer dicha separación se buscaron palabras claves de interés para este trabajo, con el fin de construir un diccionario de stopwords, las cuales fueron: foro de colaboración, foro de consulta, actividades adicionales, entrega final 1 y entrega final 2, actividades y cursos vistos.

A partir de la búsqueda de palabras claves, se observó que aquellos alumnos que hicieron las dos entregas obligatorias no sólo aprobaron la materia si no, además, hicieron un log de ingreso en todos los materiales obligatorios.

Cabe destacar que para la aplicación de learning analytics el aula virtual debe estar formateada de una forma particular para obtener información relevante para las investigaciones, por esta razón, en este trabajo no se pudo contar con cierta información de la plataforma. Esto último no permitió la construcción de variables de interés como las que se describen en los siguientes dos párrafos.

El archivo de la base de datos no tiene el log de egreso, por lo que no se sabe cuánto tiempo el estudiante permaneció conectado en la plataforma, lo que imposibilita saber si el alumno solamente descargó el contenido o si lo vio on-line en la plataforma. Dado que no se contó con el horario de egreso, no se pudo crear una variable de mucho interés en learning analytics que refiere al tiempo en que un alumno está conectado a la plataforma.

Uno de los recursos didácticos de la modalidad virtual es la explicación de algún tema a través de videos; la falta de la información de cuántos minutos el estudiante miró el video, o si completó o no la actividad de mirar el video, no está registrada en la base de datos, lo que imposibilitó crear un atributo que registre esta información.

Esta falta de información imposibilitó la creación de algunas nuevas variables que resultan importantes para este trabajo, ya que, los nuevos atributos que se pudieron generar fueron

armados solamente con el uso de los logs de entrada registrados en la base de datos, lo que hace que se pierdan algunos de los movimientos y recorridos del estudiante en el aula virtual.

A partir del conjunto de datos anonimizado entregado por las autoridades del posgrado, la minería de texto realizada y los atributos con los que se contó, junto con el resultado del rendimiento académico en la materia, se crearon los siguientes ocho atributos, a saber:

ID: identificación del alumno usando el número brindado por las personas responsables del posgrado, el rango es del 1 al 18; AO: entrega de alguna actividad adicional, codificada como 1 si entregó y 0 caso contrario; COA: comentario de retroalimentación de la actividad adicional entregada, solamente se tiene acceso a sí o no, ya que no aparece el texto de la retroalimentación por parte del docente; Co: número de conexiones totales durante el dictado de la asignatura; Fcol: número de conexiones a los foros de colaboración, este foro es para el intercambio entre pares, es decir, entre alumnos, en algunos casos simplemente es foro visto, en otros, hay colaboración; Fcon: número de conexiones a los foros de consultas, este foro es para intercambio entre alumnos y profesores, como en el foro anterior, algunos participan y otros simplemente ven el foro; NumAA: número de actividades adicionales entregadas. Solamente está el registro si entregó actividad o no, no se tiene información de la retroalimentación dada por el docente, en especial si la misma fue resuelta correctamente o no; EF: entrega final, esta variable da cuenta si el alumno hizo entrega de las dos actividades obligatorias para aprobar la materia, codificada como 1 si realizó las dos entregas correctamente y 0 en caso contrario, esta variable es la que refleja el rendimiento académico de cada alumno.

Estos ocho atributos son algunos numéricos y otros categóricos. De la base de datos resulta que ninguno de los atributos presenta datos faltantes, ni datos atípicos.

En este estudio es de especial importancia, entre las variables anteriores, la variable EF, ya que en este atributo figura el rendimiento académico de cada alumno.

Es importante notar que un alumno pudo haber visto más de una vez alguna actividad, clase, video, foro, pero esta información no está registrada en el conjunto de datos.

2.2. Indicadores. Técnicas estadísticas y data mining. Interpretación de los resultados

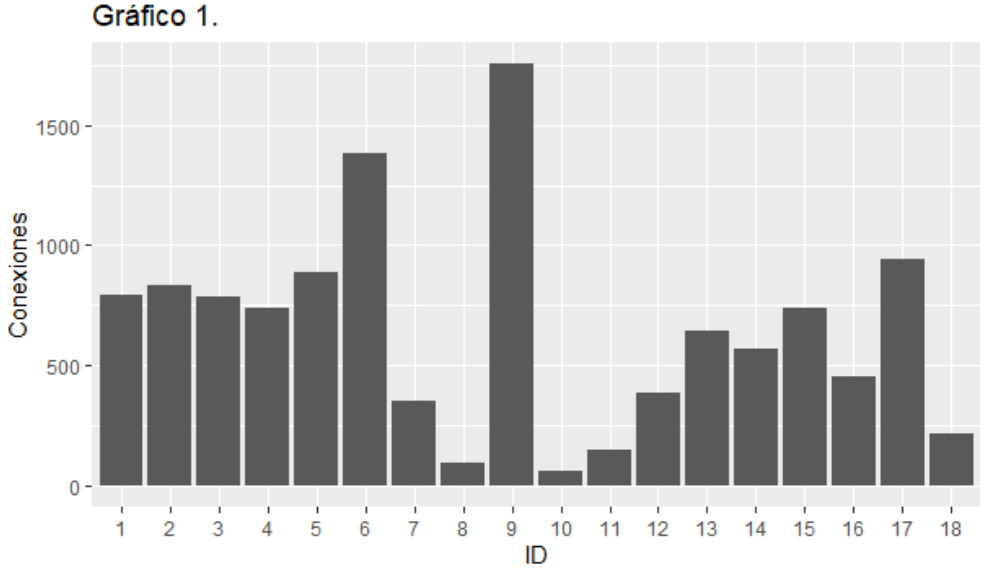
En esta sección se estudiarán y analizarán las variables definidas en la sección anterior, se propondrán indicadores del rendimiento académico junto con modelos de clusterización y se interpretarán los resultados obtenidos. Se realizará un análisis exploratorio de datos de

algunos de los atributos del conjunto de datos, para este fin, se utilizarán algunas librerías de R.

El número de conexiones o registros de los alumnos al aula virtual Moodle durante el período de cursada es de importancia en los distintos estudios de Learning Analytics (Iglesia, 2019). El atributo que refleja el número de conexiones por alumno en la base de datos está codificado con el nombre Co. Así, en Co se encuentra el número o frecuencia absoluta de conexiones por alumno, independientemente de la actividad que realice en el aula virtual Moodle.

El Gráfico 1 es un gráfico de barras que representa el número total de conexiones de cada alumno en el aula virtual. Puede observarse que los alumnos exhiben una muy diversa participación en el aula virtual.

Gráfico 1.



Fuente: elaboración propia.

Además, a partir de este gráfico se puede observar que el alumno con mayor participación con respecto a las conexiones en la plataforma es ID9, y el que tiene menor participación es ID10. El valor del atributo Co para el alumno ID9 es 1755 y este número permite construir uno de los indicadores perteneciente a una familia de indicadores que se conocen con el nombre de disparidad o brecha digital. La brecha digital permite, tanto a los profesores de la asignatura como a las autoridades del postgrado, un control de la actividad de los estudiantes en el aula virtual.

El curso constó de 23 semanas con un total de 11782 conexiones de los alumnos al aula virtual, y se obtuvo una media general de 645.55 conexiones de los alumnos. Con esta

información resulta que la media semanal de actividad de los alumnos en la plataforma es de 28.45 conexiones, este número es utilizado también para el cálculo de indicadores.

A continuación se definen qué significan y cómo se calculan algunos indicadores de disparidad o brecha digital. Se utilizan las definiciones y nombres de los indicadores como aparecen en el trabajo de Iglesia (2019).

El indicador porcentaje del total que le corresponde a cada alumno, % del total, muestra el porcentaje del total de conexiones, es decir, se obtiene haciendo para cada alumno el cociente entre el número total de conexiones realizadas por el alumno y el número total de conexiones de todos los alumnos.

El indicador media semanal por alumno, Media semanal, es el número que se obtiene como el número de conexiones del alumno dividido el número de semanas de duración del curso.

El cálculo del indicador Diferencial respecto de la media % se obtiene como la diferencia entre la media semanal de cada alumno y la media semanal de conexión de todos los alumnos.

La Disparidad respecto de la media semanal %, es un indicador que permite estudiar el comportamiento de un estudiante con respecto a todos los estudiantes, y que se construye utilizando la siguiente fórmula

(media semanal de cada alumno/media semanal de todos los alumnos).100-100

Por último, el indicador Disparidad respecto al alumno con mayor número de conexiones, en el caso de estudio resulta ser ID9, que permite estudiar el comportamiento a nivel conexiones de un estudiante con respecto al alumno con mayor número de conexiones. Este indicador se calcula mediante la diferencia de 100 con el cociente del número de conexiones totales de un alumno y el número de conexiones totales del alumno con mayor número de conexiones y este cociente multiplicado por 100.

La Tabla 1 muestra los resultados del cálculo de estos indicadores a partir del conjunto de datos.

Una visión rápida de la Tabla 1, considerando el indicador % del total, permite determinar que los cinco alumnos con mayor porcentaje de actividad son los alumnos con ID 9, 6, 17, 5 y 2, mientras que, los cinco que tienen menor porcentaje son los ID 7, 18, 11, 8 y 10.

Tabla 1. Indicadores					
ID	% del total	Media semanal	Diferencial respecto de la media %	Disparidad respecto media semanal %	Disparidad respecto al alumno ID9
1	6,75	34,57	6,12	21,51	54,7
2	7,07	36,22	7,77	27,31	52,54
3	6,67	34,17	5,72	20,11	55,21
4	6,27	32,13	3,68	12,93	57,89
5	7,51	38,48	10,03	35,25	49,57
6	11,75	60,17	31,72	111,49	21,14
7	3,01	15,43	-13,02	-45,76	79,77
8	0,81	4,13	-24,32	-85,48	94,59
9	14,9	76,3	47,85	168,19	0
10	0,49	2,52	-25,93	-91,14	96,7
11	1,26	6,48	-21,97	-77,22	91,51
12	3,26	16,7	-11,75	-41,3	78,12
13	5,46	27,96	-0,49	-1,72	63,36
14	4,82	24,7	-3,75	-13,18	67,64
15	6,28	32,17	3,72	13,08	57,83
16	3,84	19,7	-8,75	-30,76	74,19
17	7,99	40,91	12,46	43,8	46,38
18	1,86	9,52	-18,93	-66,54	87,52

Fuente: elaboración propia.

Estos indicadores calculados permiten hacer una primera agrupación de los alumnos considerando sus conexiones en el aula virtual.

Dado que el estudiante de mayor actividad en la plataforma es el ID9, si consideramos el indicador de disparidad de todos los alumnos con respecto a éste, columna 6 de la Tabla 1, se puede hacer una primera clasificación de los alumnos según el nivel de actividad en el aula virtual como activos, bastante activos, menos activos y poco activos, como se muestra en el Cuadro 1. Como este indicador se calcula referenciado al alumno con mayor actividad, ID9, la mayoría de los estudiantes se ubican en las categorías menos o poco activos, con lo que se obtiene una fuerte asimetría.

Cuadro 1. Nivel de actividad respecto de ID9.		
Nivel de actividad	Porcentaje	ID
Activos	<25%	6,9
Bastante activos	entre 25% y 50 %	5,17
Menos activos	entre 50% y 75 %	1,2,3,4,13,14,15,16
Poco activos	>75%	7,8,10,11,12,18

Fuente: elaboración propia.

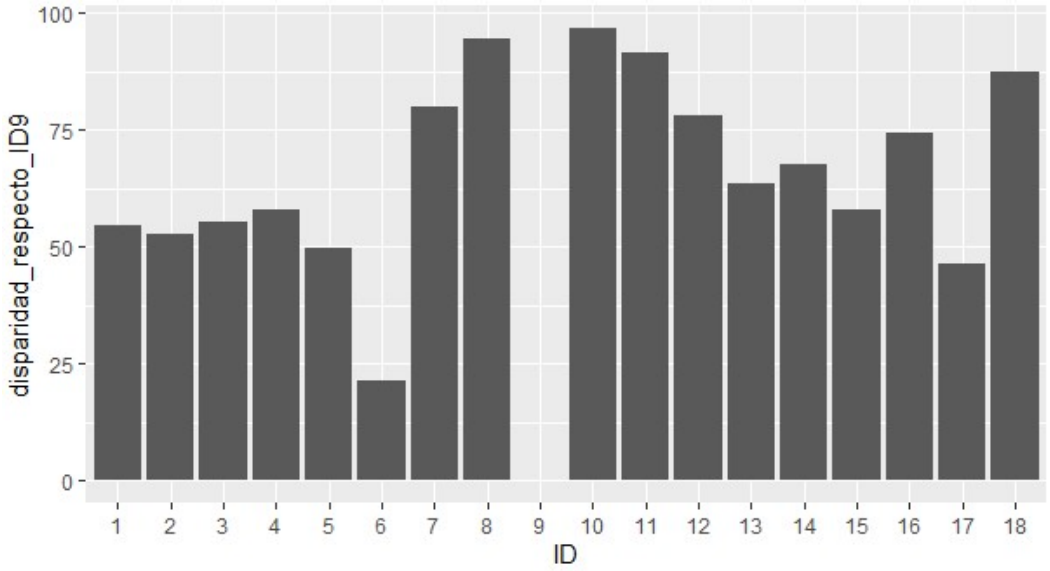
Si esta misma clasificación se realizara mediante el uso de la disparidad respecto de la media semanal %, columna 5 de la Tabla 1, la agrupación de los estudiantes en las distintas categorías se muestra en el Cuadro 2.

Cuadro 2. Nivel de actividad respecto de media semanal %.		
Nivel de actividad	Porcentaje	ID
Activos	>75%	6,9
Bastante activos	entre 25% y 75%	2,5,15,17
Menos activos	entre 25% y -25%	1,3,4,13,14
Poco activos	<-25%	7,8,10,11,12,16,18

Fuente: elaboración propia.

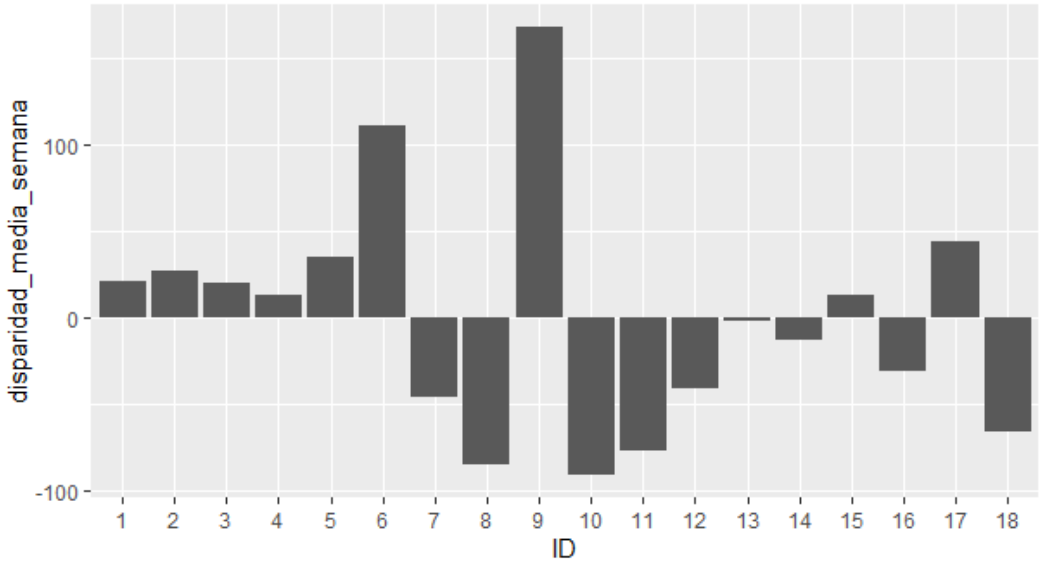
Si bien se sigue observando una asimetría en la distribución de los alumnos en las distintas categorías, ésta es un poco menor. Los siguientes gráficos, Gráfico 2 y Gráfico 3, reflejan las disparidades obtenidas a partir de la información que se encuentra en la Tabla 1.

Gráfico 2. Disparidad respecto a ID9.



Fuente: elaboración propia.

Gráfico 3. Disparidad respecto a la media semanal.



Fuente: elaboración propia.

Para identificar la brecha digital entre alumnos, se puede hacer uso de los indicadores anteriormente calculados como criterio de clasificación del número de veces que los alumnos interaccionan con el aula virtual. En el caso de estudio los indicadores fueron calculados teniendo como tiempo las semanas de duración del curso, y además fueron calculados una vez finalizado el mismo. Como mejora para cursos futuros de la materia, los indicadores se deben calcular, durante el desarrollo del curso, en períodos más cortos de tiempos, por ejemplo en vez de semanas cada dos días o cada vez que el profesor a cargo indique alguna nueva tarea a realizar, para así poder monitorear para detectar a tiempo a aquellos alumnos que presenten signos de poca actividad en el aula virtual. Esta detección de alumnos con bajo o menos nivel de actividad en la plataforma es uno de los ejes primordiales de learning analytics, y proporciona una herramienta de seguimiento de estos alumnos para evitar abandono o desaprobación de la asignatura.

La variable de interés en este trabajo es EF (Entrega Final), la cual es una variable de Bernoulli con valor 1 si se realizaron las dos entregas y ambas son aprobadas, y 0 en caso contrario, es decir, si el alumno aprueba o no la materia. Este atributo es el que refleja el desempeño académico del alumno, con la aprobación de la materia si ambas entregas fueron aprobadas.

2.3. Clusterización

En esta parte del trabajo vamos a aplicar algunas técnicas de análisis de clusterización para analizar el uso de los distintos recursos del aula virtual Moodle por parte de los alumnos.

Para llevar a cabo las agrupaciones o clusterización, se utilizan algunas de las variables numéricas que se encuentran en la base de datos con la que se cuenta. Las mismas son Co, Fcol, Fcon y NumAA. El Cuadro 3 presenta un análisis descriptivo de los 4 atributos elegidos.

Cuadro 3. Medidas resumen de las variables

	Media	Mediana	Desviación estándar	Mínimo	Máximo
Co	654.55	691	439.71	58	1755
Fcol	15.94	8.50	18.23	0	68
Fcon	22.22	11.50	20.76	0	57
NumAA	2.83	0	3.89	0	10

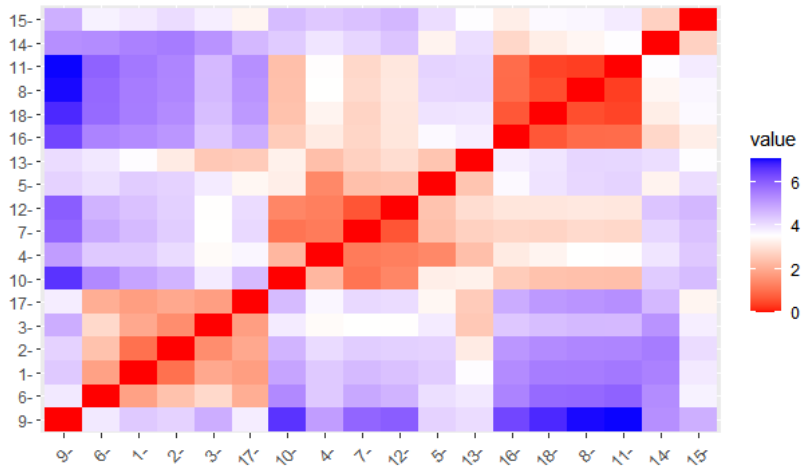
Fuente: elaboración propia.

Antes de aplicar métodos de clusterización, todas las variables fueron escaladas con media 0 y desvío estándar 1, ya que sus unidades de medición son distintas. Es de notar que también se las transformó restando a cada dato la mediana y dividiendo por la MAD (desvío absoluto mediano), como no se encontraron resultados esencialmente distintos con ambas transformaciones sólo se presentan los resultados obtenidos con las variables escaladas.

La primera técnica de minería de datos que se aplicó es clustering, que nos brinda una primera selección de atributos asociados con el uso del material por parte de los alumnos en el aula virtual, dichos atributos serán empleados para predecir la variable EF, es decir, el rendimiento académico. Esta técnica consiste en particionar las observaciones de tal manera que cada observación quede en un grupo con observaciones similares a ella y distinta a las observaciones de otros grupos formados. En este trabajo se aplica la técnica de k-medias (k-means). Realizaremos este estudio en dos etapas. En la primera etapa, mediante una matriz se decide si los datos pueden agruparse con esta técnica, y en una segunda etapa, se define el número de clústers conveniente a utilizar para formar los agrupamientos.

Para comenzar con el estudio de clusterización, se construye una matriz de distancias utilizando R, para analizar si nuestro conjunto de datos es susceptible de aplicar un análisis de clúster. El Gráfico 4 muestra la matriz obtenida, donde los colores rojo, blanco y azul representan la ausencia de relación donde azul es bajo y rojo es alto. En la misma se observan ciertas tendencias que permiten realizar agrupamientos.

Gráfico 4. Matriz de distancias.

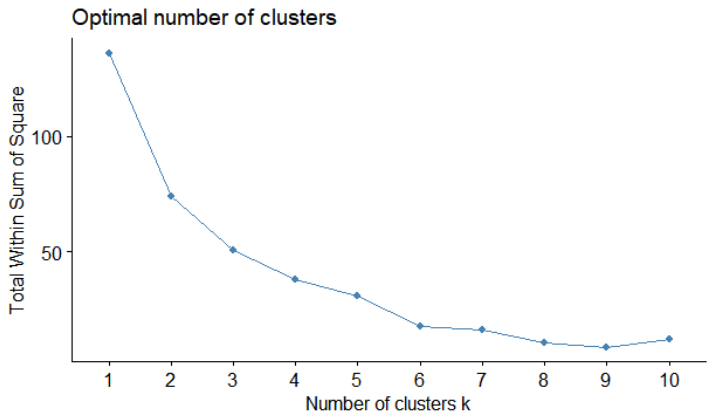


Fuente: Elaboración propia a partir de resultado obtenido con R.

Para obtener el número óptimo k de clústers se proponen tres criterios. El primero es WSS (Within Sum of Square), también conocido como el método del codo o rodilla, el cual busca minimizar la distancia media de las observaciones a su centroide. El segundo es conocido como el de la silueta (Average Silhouttes Width), este criterio evalúa cuan cerca está cada observación de un clúster a observaciones de los otros clústers. El tercero se obtiene usando una librería de R llamada NbClust, que arroja mediante la utilización de varios índices el número óptimo de clústers, este valor de k queda determinado por lo que se conoce como la regla de la mayoría.

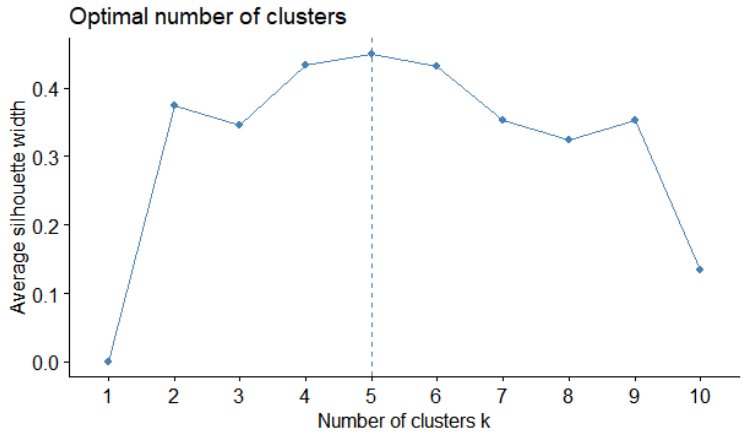
En los siguientes gráficos se muestran los resultados obtenidos a partir de ejecutar los códigos necesarios para la obtención de k con cada criterio.

Gráfico 5. Método del codo.



Fuente: Elaboración propia a partir de R.

Gráfico 6. Método de la silueta.



Fuente: Elaboración propia a partir de R.

A partir del Gráfico 5, analizando el codo o rodilla, se pueden tomar como posibles valores de k los números 3 o 4, mientras que, el Gráfico 6, el método de la silueta, propone como número óptimo k=5. Con respecto al tercer criterio, la mayoría de los índices obtenidos proponen k=3. Los Gráficos 7 y 8 muestran algunos resultados de estos índices, aquellos para los cuales R arroja su gráfico, estos son Dindex y Hubert, respectivamente.

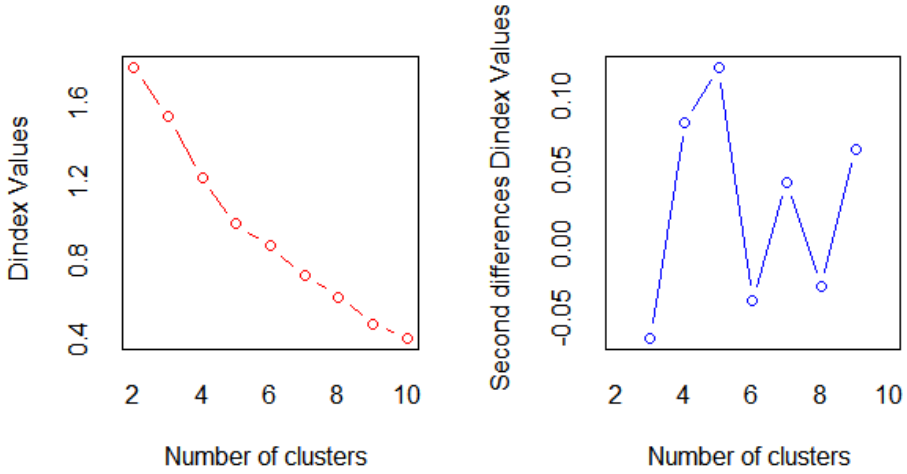


Gráfico 7. Fuente: Elaboración propia a partir de R.

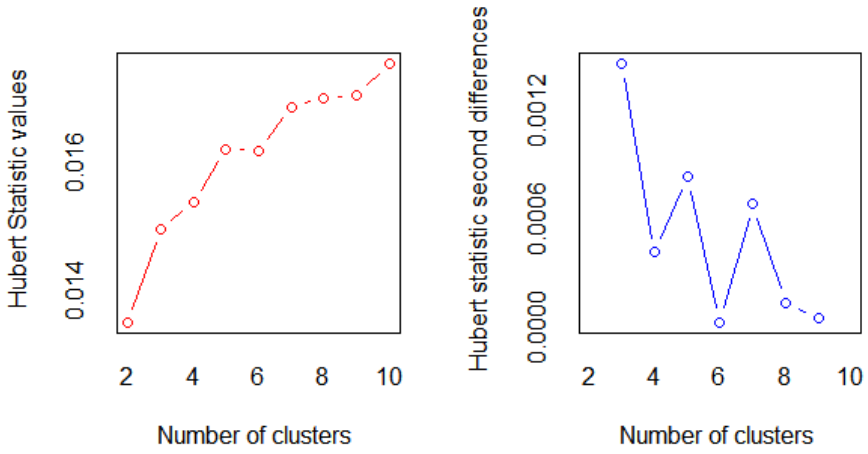
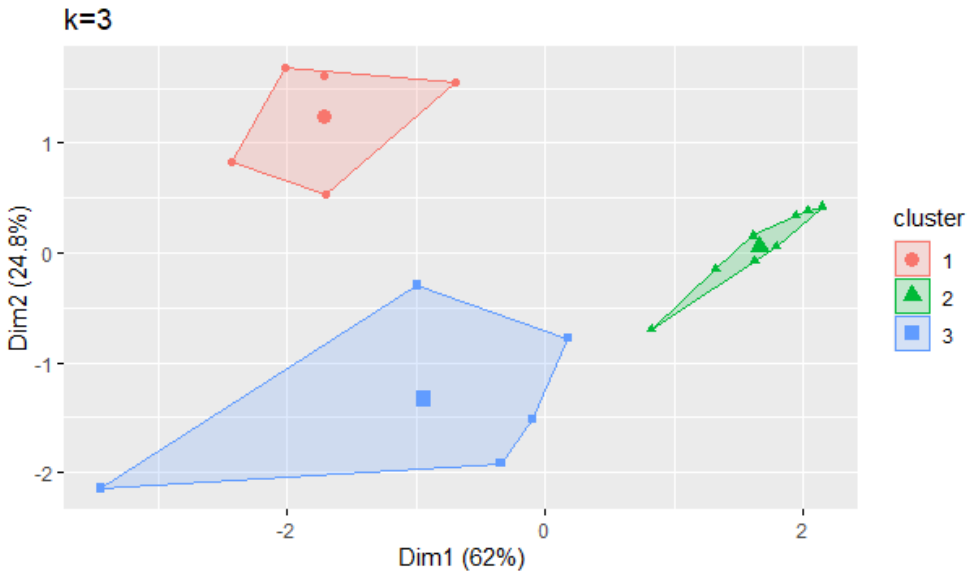


Gráfico 8. Fuente: Elaboración propia a partir de R.

Con la información recogida de los índices y las gráficos anteriores, se decidió aplicar el método k-means para el valor valores de k=3. En el Gráfico 9 se muestra la salida de R para este valor de k.

Gráfico 9. Aplicación método k-means con k=3.



Fuente: Elaboración propia

En el Cuadro 4 se muestra el agrupamiento de los 18 alumnos en cada clúster, el cual se obtiene a partir de la salida de R tras aplicar el método k-means.

Cuadro 4. Distribución de las observaciones en k=3 clústers.

Cuadro 4	Número de observación en cada clúster
Clúster 1	1 2 3 9 17
Clúster 2	4 7 8 10 11 12 16 18
Clúster 3	5 6 13 14 15

Fuente: Elaboración propia.

A partir del Cuadro 4, se pueden encontrar algunos patrones de los alumnos incluidos en cada agrupamiento. Se observa en principio que en el Clúster 2 se encuentra el 45% de los alumnos, mientras que, en los otros dos, el porcentaje de alumnos agrupados resulta igual.

A continuación, se evalúa la existencia de características similares entre los alumnos de los clústers formados. Para esto, al Cuadro 4 se agrega la variable EF, que no formó parte de la clusterización ya que es de tipo categórica.

El Cuadro 5 resume esta información, es decir, el número de alumnos por clúster según su rendimiento académico, el cual está reflejado en la variable EF.

Cuadro 5

EF	Clúster 1	Clúster 2	Clúster 3
0	0	4	2
1	5	4	3

Fuente: Elaboración propia

A partir de la información que se muestra en el Cuadro 5, en el Clúster 1 está la mayor cantidad de alumnos aprobados en la asignatura, en tanto que, en los otros dos se tiene el 50% y el 60%, respectivamente dentro de los agrupamientos.

Como cierre de este apartado, se podría extraer una conclusión de la segmentación en tres grupos con respecto a la participación de los alumnos en el aula virtual.

En el primer clúster, su participación en el foro de consulta y en el número de entregas de actividades adicionales fue alta, en este grupo todos los alumnos pudieron aprobar la materia. Cabe destacar que el alumno ID9 pertenece a este agrupamiento.

En el segundo clúster se observa que solamente el 50% aprobó la materia, y esta es su característica más importante, junto a una baja participación en foros y casi una nula cantidad de entregas adicionales.

En el tercer clúster, la participación de los alumnos es dispar respecto al uso de los foros, como también, hay alumnos con varias entregas de actividades adicionales y otros con ninguna.

3. Técnicas de predicción para el estudio del rendimiento académico

En este último apartado, con la información que se tiene en la base de datos y los análisis realizados en los apartados anteriores, se presentarán técnicas de predicción de machine learning, para el estudio del rendimiento académico de los alumnos en este posgrado virtual. Junto a esto, se indicarán las variables regresoras o predictoras de mayor peso, las cuales sirven, además, para la construcción de métricas e indicadores para la mejora del rendimiento académico de la asignatura considerada en este estudio. Del conjunto original de datos, se filtraron algunas de las variables que se definieron en el apartado 2, obteniendo de esta forma una nueva base de datos sobre la cual se aplicaron distintos modelos de regresión logística. La variable categórica EF, que plasma el rendimiento académico de los estudiantes, es la variable a predecir, mientras que las variables regresoras o explicativas, en este caso, resultan todas numéricas.

De esta manera se pretende cumplir con el objetivo planteado para obtener un modelo predictor del rendimiento académico a partir de las variables con las que se dispone.

3.1. Regresión Logística

La técnica de machine learning que se aplicó es regresión logística (Peña, 2002). Esta, como otras técnicas predictivas, utiliza todas o algunas de las variables de la base de datos, con el objeto de predecir valores desconocidos de otra variable. El modelo de regresión logística se utiliza para modelar respuestas dicotómicas (presencia o ausencia de una condición) en función de un conjunto de variables (covariables) que posiblemente afecten la respuesta. La regresión logística provee una estrategia de modelización general y de interpretación directa. También permite que las covariables sean cualitativas, ordinales o cuantitativas. Desde el punto de vista teórico, el modelo de regresión logística forma parte de una familia de modelos llamados Modelos Lineales Generalizados. En estos modelos se modela una transformación de la media de la variable respuesta como una combinación lineal de las variables explicativas. En el modelo de regresión logística la variable respuesta es una variable Bernoulli, que toma solo dos posibles valores 1 (éxito) y 0 (fracaso).

En nuestro estudio, la variable a predecir es EF, que como se explicitó anteriormente, refleja el rendimiento académico. Esta variable tiene dos categorías 0 y 1, 0 indica que el alumno no aprobó la materia y 1 que lo hizo.

Las variables explicativas o regresoras utilizadas son numéricas y resultaron ser las siguientes: Co, Fcol; Fcon y NumAA, las cuales como en el apartado anterior fueron estandarizadas para poder aplicar el método.

En este trabajo, se ajustó un modelo de regresión logística simple para predecir EF a partir de cada una de las variables regresoras.

Cabe destacar que el número de observaciones que se dispone en este conjunto de datos es bajo, 18, pero los métodos de predicción aplicados pueden ser replicados para muestras más grandes. Por esto, se decidió realizar la regresión logística sin partir el conjunto de datos en grupo de entrenamiento y testeo. A partir de esta decisión se trata de identificar qué variables regresoras resultan mejores para predecir el rendimiento académico, con el objetivo de construir un modelo.

Para definir la base de datos con la cual trabajar se consideraron las variables EF, variable a predecir, y las variables numéricas Co, Fcol y Fcon, Se descartaron atributos que por tener una gran cantidad de ceros no aportaban a la predicción.

Definida esta base, se aplicaron en R diferentes modelos de regresión logística cambiando las variables regresoras, para así obtener un modelo que mejor clasifique el rendimiento académico a partir de ciertos comportamientos de los alumnos en el aula virtual. En el siguiente subapartado se definirán las métricas que se utilizarán para evaluar los modelos.

3.2. Métricas

Cuando la variable de respuesta es binaria, como lo es la variable a predecir en este trabajo rendimiento académico, hay dos posibles valores reales y dos posibles valores de predicción o predichos. A partir de estas opciones podemos crear lo que se conoce como Matriz de Confusión. Esta matriz es de dimensión 2x2 y es de la siguiente forma

		Valor Real	
		Positivo	Negativo
Valor Predicho	Positivo	Verdadero Positivo	Falso Positivo
	Negativo	Falso Negativo	Verdadero Negativo

Fuente: elaboración propia.

Donde, Verdadero Positivo (VP) significa que el valor real es positivo y la prueba predice un positivo; Verdadero Negativo (VN) significa que el valor real es negativo y la prueba predice un negativo; Falso Negativo (FN) refiere a que el valor real es positivo y la prueba predice un negativo y por último, Falso Positivo (FP) que refiere a que el valor real es negativo y la prueba predice un positivo. Observar que en la diagonal principal se ubican el número de casos que fueron correctamente clasificados.

A partir de esta Matriz de Confusión se definen las siguientes métricas o medidas de evaluación de un modelo (Peña, 2002).

En primer lugar definimos la sensibilidad o Recall, que es la probabilidad de obtener un verdadero positivo, su cálculo se obtiene haciendo el cociente entre VP y la suma de VP más FN. Esta métrica da información sobre el desempeño de un clasificador con respecto a los falsos negativos, es decir, cuanto falla el modelo.

En segundo lugar tenemos la especificidad, que es la probabilidad de obtener un falso positivo. Su cálculo se realiza como el cociente entre FP y la suma de FP con VN.

Para tener información sobre el desempeño del clasificador respecto a los falsos positivos, se tiene la tercera métrica, a saber, la precisión o Precision, que se calcula como el cociente entre VP y la suma de VP con FP.

Por último, la cuarta métrica considerada es Accuracy, que es la proporción total de casos correctamente clasificados, para su cálculo se realiza el cociente entre VP+VN y VP+FP+VN+FN.

Cabe destacar que no hay una métrica que sea mejor que las otras. Cada una proporciona información útil sobre el rendimiento del modelo en diferentes aspectos del problema de clasificación.

3.3. Modelos de Regresión Logística

Se corrieron en R siete modelos de regresión logística. En todos los casos EF fue el atributo a predecir, en tres de ellos se consideró la regresión con respecto a cada una de las variables regresoras de la base (Co, Fcol y Fcon), en tres tomándolas de a pares para predecir EF y en una las tres variables en conjunto. En los siete modelos se obtuvo la Matriz de Confusión y cada una de las métricas correspondientes.

En la Tabla 2 se pueden observar los resultados obtenidos con R cuando se consideró las variables EF y Co, a esta regresión logística la llamamos Modelo 1.

Tabla 2. Matriz de Confusión entre EF y Co		
	True 1	True 0
Pred 1	9	1
Pred 0	3	5

Fuente: elaboración propia según resultados obtenidos con R.

La suma de los elementos en la diagonal de la Matriz de Confusión indica el número de observaciones que fueron clasificadas correctamente, que resultan ser 14. Hay una observación que fue clasificada como que aprobó la materia cuando en la realidad no lo hizo, y tres observaciones que fueron clasificadas como desaprobadas y en la realidad estaban aprobadas, que resultan ser falsos positivos y falsos negativos, respectivamente. Cuando una observación sea clasificada como perteneciente al grupo 1 vamos a decir que es positiva, es decir, el alumno aprobó la materia.

La siguiente Tabla 3 muestra los resultados de cada métrica calculadas a partir de la predicción.

Tabla 3. Métricas.		
Accuracy	Precision	Recall
77,77 %	90%	75%

Fuente: elaboración propia según resultados obtenidos con R.

La performance del modelo, la métrica Accuracy del 77,77%, sugiere que es aceptable la elección del modelo con variable regresora Co.

Trabajando de manera análoga, se obtuvieron las Matrices de Confusión y las métricas correspondientes para lo que se llamó Modelo 2 (EF y Fcol), Modelo 3 (EF y Fcon), Modelo 4 (EF, Co y Fcol), Modelo 5 (EF, Co y Fcon), Modelo 6 (EF, Fcol y Fcon) y Modelo 7 (EF, Co, Fcon y Fcol).

La siguiente Tabla 4 muestra las métricas, de cada uno de los modelos anteriores, obtenidas a partir de los resultados arrojados por R.

Tabla 4. Métricas de los modelos.			
Modelo	Accuracy	Precision	Recall
Modelo 1	77,77%	90%	75%
Modelo 2	61,11%	77,77%	58,33%
Modelo 3	66%	80%	66,66%
Modelo 4	83,33%	84,61%	91,66%
Modelo 5	88,88%	92%	83,33%
Modelo 6	55,55%	83,33%	41,66%
Modelo 7	88,88%	68,75%	91,66%

Fuente: elaboración propia según resultados obtenidos con R.

De la Tabla 5, se observa que los modelos con mejor performance son los modelos que tuvieron mayor Accuracy, aunque no tiene por qué ser la mejor métrica, es decir, Modelo 4, Modelo 5 y Modelo 7. En este trabajo se decidió considerar aquellos modelos, que además de tener Accuracy alta, tienen Recall o Sensibilidad alta, esto es para evitar falsos negativos. Los Modelos 5 y 7 son los que tienen mayor valor en las métricas Accuracy y Recall.

Para finalizar este apartado, teniendo en cuenta las métricas obtenidas para cada modelo de regresión logística considerado, se tomará como modelo para realizar la regresión del rendimiento académico al Modelo 5. Las variables regresoras para este modelo son Co y Fcons. Las tres métricas para este modelo resultaron ser buenas.

Hasta aquí los distintos tipos de Regresiones Logísticas que se realizaron con la base considerada. Cabe mencionar que el Modelo 5 considerado es el modelo elegido. Si se contara con más registros, mayor número de estudiantes que participen en la materia, se podría realizar una partición del conjunto de datos en grupo de entrenamiento y grupo de control para analizar la confiabilidad del modelo propuesto.

Conclusión

En la introducción de este trabajo se expuso el objetivo principal que fue la construcción de indicadores y modelos predictivos del rendimiento académico, utilizando algunas técnicas de learning analytics, para la detección de patrones relacionados con la interacción de los alumnos en el aula virtual de una asignatura de posgrado.

Para la elaboración de este trabajo se contó con una base de datos de los alumnos de una asignatura de un posgrado virtual, la misma fue entregada en forma anonimizada por las autoridades del posgrado. La base fue obtenida a partir de los registros de las interacciones de los alumnos con el aula virtual de la materia en la plataforma Moodle en un cuatrimestre de 23 semanas.

La actividad de procesamiento de los datos llevado a cabo en el primer apartado es vital para poder conseguir una base de datos de calidad y atributos sustanciales para ser utilizados en los otros apartados, tanto para técnicas de machine learning como también para conseguir una menor complejidad y mejor calidad en los modelos predictivos.

Se pudo constatar a través del desarrollo del trabajo, que las técnicas de clusterización pudieron ser aplicadas de una manera simple para la obtención de atributos que caractericen a los agrupamientos. Lo mismo fue observado en las técnicas de clasificación, obteniendo altos niveles de confiabilidad de los algoritmos propuestos, con el objetivo de predecir el rendimiento académico de los estudiantes.

En el trabajo se tuvieron en cuenta solamente del conjunto de atributos disponibles aquellos que estaban relacionadas con el uso del aula virtual y los distintos recorridos de los estudiantes en la misma. Cabe destacar que el rendimiento académico de un estudiante no depende exclusivamente de las variables consideradas en este trabajo, para realizar un trabajo más detallado de este rendimiento, se debería disponer de información referente a factores cognitivos, sociales, demográficos, personales, entre otros. Para poder llevar a cabo esto se requiere una configuración especial de la plataforma Moodle que permita obtener la información de interés.

La información y resultados arrojados en el segundo apartado, con las técnicas de agrupamiento, evidenciaron que un nivel bajo en el uso de los recursos de la plataforma, principalmente lo referente a entregas adicionales y utilización de foros de comunicación,

llevaron a no aprobar la materia, mientras que, aquellos alumnos que tuvieron una mayor participación en estos atributos alcanzaron un mejor nivel de rendimiento.

Los indicadores construidos en el apartado 1, Tabla 1, han permitido una primera clasificación del rendimiento académico de los estudiantes, como quedó reflejado en los cuadros 1 y 2. Durante el desarrollo del curso virtual, estos indicadores se pueden ir midiendo y actualizando de forma semanal, de manera tal que autoridades y profesores dispongan de una medida para poder detectar de forma temprana a aquellos estudiantes que puedan estar en las categorías menos activos y poco activos. Además, con estos indicadores se pueden detectar alumnos que necesitan un mayor seguimiento, tratando de evitar de esta manera que desapruebe o abandone la materia, esto último es uno de los ejes primordiales en learning analytics.

Los modelos de regresión logística desarrollados en el apartado 3, muestran con alta precisión, que un control sobre las comunicaciones en foros de consultas y el número de conexiones con la plataforma debe ser imperioso. Desde lo pedagógico, la cátedra de la materia deber prestar atención a estos dos atributos, foros de consultas y conexiones con la plataforma, promoviendo su uso y sirvan para construir, por ejemplo, una parte de la nota que forma la nota final de la materia.

Como ya se dijo en este trabajo se utilizaron los datos obtenidos en un cuatrimestre de una materia de un posgrado virtual. Es esperable que el trabajo contribuya no solamente a la cátedra que generó los datos, sino, además, en el futuro a los otros cursos que son parte del plan de estudio. De esta forma, se incrementa el volumen de datos en la base y se pueden obtener mejores precisiones en los modelos predictivos, como también, una mayor generalización de los resultados.

Para un futuro trabajo es importante que, en la plataforma, además de los atributos que se obtuvieron, se pueda tener acceso a algunos datos personales, encuestas de satisfacción, horario de egreso, actividades realizadas sobre la plataforma, entre otras.

Se espera que resulte de utilidad para las autoridades del posgrado virtual el presente trabajo, permitiendo a las mismas guiar a los alumnos y docentes, al disponer de indicadores y de una regresión logística, la obtención de una mejora en la evolución de la enseñanza y aprendizaje.

Referencias bibliográficas

Agesic. (2017). *Guía de disociación y anonimización de datos personales.* Montevideo.

Aldas, J., & Uriel, E. (2017). *Análisis Multivariado aplicado con R.* Madrid: Paraninfo.

Anderson, T. (2003). *An introduction to multivariate statistical analysis.* New Jersey: Willey.

Appelbaum, D., Kogan, A., & Vasarhelyi, M. (2017). *Big Data and Analytics in the Modern Audit Engagement.* American Accounting Association.

Arnold, K., & Pistilli, M. (2012). Course signals at Purdue. Using learning analytics to increase student success. *2nd International Conference on Learning Analytics and Knowlegde.* Vancuouver: ACM.

Crowley, M. (2012). *The R book.* Wiley.

Dalgaord, P. (2008). *Introductory Statistics with R.* USA: Springer.

Dehghantanha, A., & Choo, K. (2019). *Handbook of big data and IoT security.* Springer.

Dietz-Uhler, B., & Hurn, J. (2013). Using Learning Analytics to Predict (and Improve) Student Succes: A Faculty Perspective. *Journal of Interactive Online Learning*, 17-26.

Floric, E., Igual, L., & Puertas, E. (2020). Learning Analytics para el personal académico. *Actas de la Jenui* (págs. 245-252). Granada: Universidad de Granada.

García Herrero, J., & al., e. (2018). Ciencia de datos: Técnicas analíticas y aprendizaje estadístico en un enfoque práctico. . *Alfaomega*.

García Pérez, A. (2005). *Métodos avanzados de estadística aplicada. Técnicas avanzadas.* Madrid: UNED.

Iglesia Villasol, M. (2019). Learning Analytics para una visión tipificada del aprendizaje de los estudiantes. Un estudio de caso. *Revista Iberoamericana de Educación* , 55-87.

Jhonson, D. (2000). *Métodos multivariados aplicados al análisis de datos.* México: Thomson.

Lu, O., Huang, J., & al., e. (2018). Applying learning analytics for the early prediction of students´ academic performance in blended learning. *Educational Technology and Society*, 220-232.

Moodle. (s.f.). Obtenido de https://moodle.org/?lang=es

N°25.326, L. (s.f.). *InfoLEG*. Obtenido de http://servicios.infoleg.gob.ar/infolegInternet/anexos/60000-64999/64790/norma.htm

Peña, D. (2002). *Análisis de datos multivariante.* Madrid: McGraw Hill Interamericana.

Picciano, A. (2012). The Evolution of Big Data and Learning Analytics in American Higher Education. *Journal of Asynchronous Learning Network*, 9-20.

RapidMiner. (s.f.). Obtenido de https://rapidminer.com/

Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. (2018). *Introduction to Data Mining.* USA: Pearson Education.

The R Project for Statistical Computing. (s.f.). Obtenido de https://www.r-project.org/

Travieso, J., & Moreno, M. (2006). *La protección de los datos personales y de los sensibles en la ley 25.326.* Buenos Aires: La Ley.

Urbina-Nàjera, A. (2021). Variables que influyen en el rendimiento de los estudiantes de posgrado: Una perspectiva desde la analítica del aprendizaje. *TELOS*, 26-50.

Van-Barneveld, A., & Arnold, K. (2012). Analytics in Higher Education: Establishing a common language. *EDUCAUSE*, 1-11.

Yu, T., & Jo, I. (2014). Educational technology approach toward learning analytics: relationship between student online behaviour and learning performance in higher education. *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, 70-81.

Reporte Trabajo Final Integrador de la Especialización de Gabriel Duarte
"Aplicación de learning analytics a un curso de posgrado virtual de una universidad pública"
Tutora: Patricia Beatriz Girimonte
Fecha: Septiembre 2023

El presente trabajo tiene como objetivo construir indicadores y modelos predictivos del rendimiento académico de los estudiantes de una asignatura de un posgrado en modalidad virtual de una universidad pública, aplicando técnicas de learning analytics.
Considero que el objetivo planteado y su desarrollo es relevante dado que permite detectar patrones de comportamiento relacionados con la interacción de los alumnos en un aula virtual, lo cual sería de utilidad para la mejora de la enseñanza y la toma de decisiones tanto a nivel académico, administrativo o de gestión.
El problema está identificado y definido correctamente. Su desarrollo en tres apartados es pertinente. En el primer apartado, se describen, citando la bibliografía correspondiente, los conceptos de learning analytics, su importancia y alcance. En este primer apartado, si bien los datos analizados fueron recibidos en forma anonimizada, se hace referencia a la importancia de la anonimización cuando se trabaja con datos personales, y en particular a la ley 25.326 vigente en nuestro país.
En el segundo apartado se aplican correctamente las técnicas de learning analytics, para luego realizar un análisis de cluster con el objeto de construir un indicador del rendimiento académico. En el tercer apartado se utiliza el modelo de regresión logística para predecir la aprobación o no del curso.
Si bien el trabajo no explora todas las variables que podrían ser de interés para la predicción del rendimiento académico, dado que, como es mencionado, no se contó con esa información se obtienen buenos resultados que permitirán abordar trabajos futuros en los que puedan ser analizadas otros variables predictoras.
El planteo del problema, los objetivos y la hipótesis se encuentran articulados con la formación de grado, la carrera profesional del autor y con la especialización, dado que los conceptos abordados se encuentran dentro de los contenidos de los módulos de la especialización.