RECIE
Revista Electrónica Científica de Investigación Educativa

Volumen 5, núm. 2, enero-diciembre 2021
ISSN: 2594-200X

Obtención de un modelo de *Learning Analytics* con información de un LMS

Obtaining a Learning Analytics model with information from an LMS

Leonardo Nevárez Chávez • Marisela Ivette Caldera Franco • Gregorio Ronquillo Máynez

Leonardo Nevárez Chávez. Tecnológico Nacional de México, campus Chihuahua II, México. Es docente del programa de maestría en Sistemas Computacionales. Cuenta con estudios como maestro en Ciencias en Cómputo Aplicado. Miembro del Cuerpo Académico Educación Matemática y Computación ITCHID-CA-2, LIIADT: Didáctica de la Matemática/Tecnología Aplicada a la Educación. Correo electrónico: leonardo.nc@chihuahua2.tecnm.mx. ORCID: https://orcid.org/0000-0003-2857-5124.

Marisela Ivette Caldera Franco. Tecnológico Nacional de México, campus Chihuahua II, México. Es docente del programa de maestría en Sistemas Computacionales. Doctora en Educación. Miembro del Cuerpo Académico Educación Matemática y Computación ITCHID-CA-2, LIIADT: Didáctica de la Matemática/Tecnología Aplicada a la Educación. Perfil Deseable Prodep. Correo electrónico: marisela.cf@chihuahua2.tecnm.mx. ORCID: https://orcid.org/0000-0001-5574-5817.

Gregorio Ronquillo Máynez. Tecnológico Nacional de México, campus Chihuahua II, México. Es docente en la maestría en Sistemas Computacionales y en licenciatura. Doctor en

Resumen

En las instituciones de educación superior, como resultado del proceso de enseñanza y aprendizaje, se generan y obtienen datos en diferentes sistemas computacionales, tales como sistemas internos, LMS (*Learning Management System*), redes sociales, entre otros. La información que se obtiene de estos sistemas raramente es utilizada para su análisis, retroalimentación y mejora de los procesos. Considerando solamente un LMS, estos producen información importante, como accesos de los estudiantes, secciones o elementos vistos, entrega de tareas y su cumplimiento a plazos, así como la participación en foros y otras actividades. Esta investigación tiene como propósito describir el uso de la información contenida en la bitácora, la cual es generada por el LMS Moodle, con el objetivo de establecer un modelo de *Learning Analytics* que lleve a predecir el rendimiento de los estudiantes. El modelo es incorporado en una aplicación informática mediante una interfaz amigable, y sirve para dar a conocer los resultados del modelo a tutores, profesores o personal autorizado. La aplicación identifica estudiantes en riesgo académico y se sugiere utilizarla como apoyo para decidir una posible intervención académica en beneficio de los estudiantes. Se usan técnicas de regresión simple, regresión múltiple, agrupamiento y análisis de componentes principales para obtener diferentes modelos de predicción. Se aplica la metodología CRISP-DM como guía en el proceso de desarrollo.

Palabras clave: Análisis de datos, rendimiento académico, plataforma Moodle, modelos de análisis, intervención educativa.

Abstract

As a result of the teaching and learning process, data in higher education institutions is generated and obtained in different computational systems, such as internal systems, the LMS (Learning Management System), social media, among others. The information obtained from these systems is rarely used for its analysis, feedback, and processes improvement. Considering only one LMS, these produce important information, as students' logins, sections or seen elements, delivery of homework and its compliance in time limits, as well as the participation in forums and other activities.

Recibido: 30 de mayo de 2021 | Aprobado: 22 de noviembre de 2021 | Publicado: 10 de diciembre de 2021
DOI: https://doi.org/10.33010/recie.v5i2.1314

Revista Electrónica Científica de Investigación Educativa
Volumen 5, núm. 2, enero-diciembre 2021, pp. 313-333

Administración y maestro en Sistemas de Información por la Facultad de Contaduría y Administración de la Universidad Autónoma de Chihuahua, e ingeniero en Sistemas Computacionales por el Instituto Tecnológico de Chihuahua II. Correo electrónico: gregorio.rm@chihuahua2.tecnm.mx. ORCID: https://orcid.org/0000-0002-8939-5736.

This research has the purpose of describing the use of the information contained in the logs, which is generated by the Moodle LMS, with the objective of establishing a Learning Analytics model, leading to the prediction of the students' performance. The model is incorporated in an informatic application through a friendly interface, and it serves to make known the model results to tutors, professors or authorized staff. The application identifies students in academic danger, and it is suggested to be used as a support to decide a possible academic intervention on behalf of students. Simple regression, multiple regression, grouping, and main components analysis techniques are used to obtain different prediction models. The CRISP-DM methodology is applied as a guidance in the development process.

Keywords: Data analysis, academic performance, Moodle platform, analysis models, educational intervention.

Introducción

Actualmente, en el Tecnológico Nacional de México (TecNM), campus Chihuahua II, no se cuenta con sistemas *Learning Analytics* (LA) que sirvan de apoyo al proceso educativo.

En el entorno internacional, diferentes centros de educación utilizan LA como herramienta de apoyo en sus actividades. Sclater, Peasgood y Mullan (2016) identificaron en un mapa diversas universidades y mencionan actividades académicas que con esta herramienta se han desarrollado.

En la literatura acreditada se encuentran ejemplos de la gestión, desarrollo y uso del LA. Se mencionan a continuación algunos de estos. Daud, Aljohani, Abbasi, Lytras, Abbas y Alowibdi (2017) establecieron la predicción del rendimiento de los estudiantes mediante técnicas de LA, para ello usaron sus datos sociales y económicos. El objetivo principal fue determinar si completarían con éxito su curso. En general los datos considerados están en las categorías de gasto e ingreso familiar, información general y bienes familiares.

Viberg, Hatakka, Bälter y Mavroudi (2018) presentaron un estudio del estado actual de LA en educación superior, sustentado en 252 publicaciones de este tema durante el periodo 2012 al 2018. La pregunta principal que intentaron responder es relacionada con el conocimiento científico actual acerca de la aplicación de LA en educación superior. En este sentido identificaron los enfoques de las investigaciones, los métodos y evidencias de LA.

Martínez y Moreno-Ger (2018) realizaron una comparación de algoritmos de clústeres para LA, sobre conjuntos de datos educativos, encontraron que los algoritmos K-means y PAM tuvieron el mejor desempeño en la categoría de partición, en tanto el algoritmo DIANA lo obtuvo entre los algoritmos jerárquicos.

Respecto al LA, utilizando un LMS, se encuentra el estudio de Simanca, Herrera, Crespo, Baena y Burgos (2019), en este describen el desarrollo de un sistema en el que se utilizan técnicas de LA a través de información que se obtiene de un LMS. El

sistema desarrollado fue usado para ejecutar una estimación de aquello que puede ocurrir con un estudiante y con ello realizar un seguimiento con el objetivo de que culmine con éxito el curso. Mencionan también la información obtenida del LMS para su estudio: número de inicios de sesión, tiempo total de conexión, rendimiento individual contra grupal, calificaciones de actividades y uso de recursos (como lecturas y envíos).

Dawson, Joksimovic, Poquet y Siemens (2019) plantean la manera de incrementar el impacto del LA. Realizaron un estudio de publicaciones en LA durante los años 2011 al 2018, en el cual analizan aspectos como: enfoque del estudio, tipos de datos usados, propósito, marco institucional, así como la escala de la investigación y su aplicación. Concluyeron que la mayoría de esfuerzos han sucedido a una escala pequeña y centrados en aspectos de tecnología. Enfatizan abordar los trabajos de LA con un enfoque multi-disciplinario, tal como lo requiere la complejidad de la educación.

Herodotou, Rienties, Boroowa, Zdrahal y Hlosta (2019) describen la aplicación a gran escala del LA predictivo (PLA) en la educación superior, coleccionando para ello información de diferentes fuentes. El sistema fue aplicado para identificar estudiantes en riesgo durante los cursos de educación a distancia.

Gasevic, Tsai, Dawson y Pardo (2019) diseñaron una propuesta para iniciar la adopción institucional de un proyecto de LA. En su estudio identificaron los retos críticos que requieren atención, de manera que con el LA se logre un impacto de largo término en la investigación y práctica del aprendizaje, así como en la enseñanza.

Respecto a las desventajas e inconvenientes que se pueden presentar con LA, Selwyn (2019) presentó un estudio crítico del tema. Mencionando que el LA debe integrar los aspectos social, cultural, político y económico. Destaca la necesidad de contar con aplicaciones de LA abiertas y accesibles, las cuales proporcionen un control y supervisión genuino a sus usuarios y reflejen la realidad vivida por los estudiantes.

En el artículo de Romero y Ventura (2020) se aborda un estudio del estado del arte en los temas de *Educational data mining* (EDM) y LA. Revisaron las principales publicaciones, hitos clave, el ciclo de descubrimiento del conocimiento, principales entornos educativos, herramientas específicas, conjuntos de datos gratuitos disponibles, métodos utilizados, principales objetivos, incluyendo las tendencias futuras.

La *analítica del aprendizaje* (LA) es la medición, recopilación, análisis y reporte de datos sobre los alumnos y sus contextos, con el fin de comprender y optimizar el aprendizaje y los entornos en los que se produce (Lang *et al.*, 2017, citados en Romero y Ventura, 2020).

"El crecimiento del aprendizaje en línea, particularmente en la educación superior, ha contribuido al avance del LA, puesto que es posible capturar los datos de los estudiantes y disponerlos para su análisis" ("Analítica del aprendizaje", 2022).

En diferentes entornos se usa el concepto de *Blended Learning*, en el cual la educación presencial se apoya utilizando tecnologías como los LMS. Estos sistemas registran información importante sobre el comportamiento de los usuarios.

Revista Electrónica Científica de Investigación Educativa
Volumen 5, núm. 2, enero-diciembre 2021, pp. 313-333

Además de los LMS, los alumnos usan redes sociales, herramientas en línea, sistemas internos en las instituciones educativas, bibliotecas digitales, etc. Estas aplicaciones guardan datos, como las elecciones que realizan, clics, patrones de navegación, tiempo en realizar tareas, desarrollo de conceptos en foros y discusiones, entre otros.

En las instituciones educativas, como el TecNM y sus campus, se cuenta con datos e información generada por los alumnos, que aún no se ha utilizado para propósitos de investigación a través de LA, la cual es posible utilizar como medio de apoyo en el proceso de enseñanza-aprendizaje.

Metodología

Considerando a los LA como un área especializada de ciencia de datos, se eligió la metodología CRISP-DM (Cross-industry standard process for data mining) como la adecuada para este proyecto, por ser en la actualidad una de las más ampliamente usadas. Las etapas generales se muestran en la figura 1.

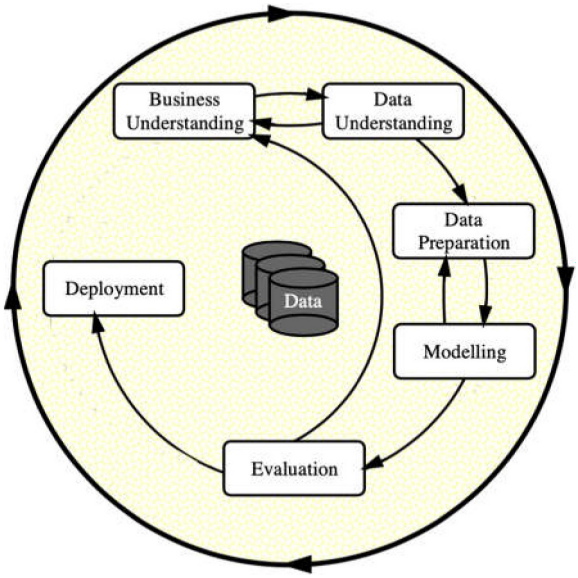

Figura 1. Fases del modelo CRISP-DM.
Fuente: Wirth y Hipp (2000).

En la figura 1 se observa el ciclo externo que indica la naturaleza cíclica en los procesos de ciencia de datos. Se describen enseguida las actividades específicas del proyecto aplicando la metodología CRISP-DM.

Comprensión del negocio

En principio, se obtuvo la misión y visión del TecNM y del TecNM campus Chihuahua II. Se analizó el perfil de egreso de un ingeniero en Sistemas Computacionales (ISC),

puesto que en la etapa inicial del proyecto se consideran estudiantes de esta carrera. Se determinaron, además, los objetivos del negocio.

El objetivo es desarrollar un modelo de Learning Analytics que lleve a predecir el rendimiento de los alumnos, sustentado en la bitácora del LMS Moodle, de modo que, por medio de una aplicación informática, sea posible detectar a los estudiantes en riesgo académico.

Comprensión de los datos

Se obtuvieron las bitácoras del LMS Moodle de los años 2019 y 2020 (del 21 de agosto del 2019 al 20 de agosto del 2020), los cuales corresponden a datos del segundo semestre del 2019 y primer semestre del 2020. El sistema generó un archivo tipo CSV (formato de archivo de texto separado por comas). La información obtenida contiene alrededor de cuatro millones de renglones o registros (exactamente fueron 3,999,427).

La bitácora contiene los campos «Hora», «Nombre completo del usuario», «Usuario afectado», «Contexto del evento», «Componente», «Nombre del evento», «Descripción», «Origen» y «Dirección IP». Se puede ver una muestra de datos reales en la tabla 1. Con propósito descriptivo solo se muestran dos registros de la bitácora. La bitácora original tiene información de todos los cursos que usan Moodle en el periodo indicado. Para este estudio se usan solamente registros de la bitácora que tienen relación con la materia de "Programación web", pues de esta se obtuvo la información de calificación final de los estudiantes.

Tabla 1. Bitácora del LMS.

Hora	Nombre completo usuario	Usuario afectado	Contexto del evento	Compo-nente	Nombre del evento	Descripción	Origen	Dirección IP
20/08/20, 13:45			Curso: Análisis crítico de la arquitectura y el arte IV- A.M.	Sistema	Curso visto	The user with id '5377' viewed the course with id '833'	Web	177.236.40.24
20/08/20, 13:45			Curso: Taller Inv. I ago.-dic. 2020	Sistema	Curso visto	The user with id '3661' viewed the course with id '824'	Web	189.237.216.7

Fuente: Elaboración de los autores (se omite información del usuario).

El aspecto mencionado delimita entonces el alcance de este estudio a una sola materia durante dos semestres (segundo del 2019 y primero del 2020), por lo cual la parte de análisis, modelado y predicción tendrán esta limitación. Para obtener modelos y predicciones más robustas se debe ampliar el número de cursos y de semestres.

En particular el campo "Nombre del evento" indica la acción concreta que el usuario realizó en el LMS y será fundamental para el análisis de la información. Se identifican 222 nombres de eventos diferentes. Los nombres de eventos más comunes se obtienen y muestran en la tabla 2.

Revista Electrónica Científica de Investigación Educativa
Volumen 5, núm. 2, enero-diciembre 2021, pp. 313-333

Tabla 2. Nombres de eventos más comunes.

Nombre evento	Ocurrencias
Curso visto	926713
Módulo de curso visto	832531
El estatus del envío ha sido visto	336064
El usuario ha ingresado	269266
Intento de examen visto	249214
Calificación eliminada	129557
Formato de envío visto	101104
Notificación enviada	96033
Usuario calificado	92694
Falló el ingreso del usuario	75661
Se ha enviado un envío	53792
Se ha actualizado un archivo	53378
Finalización de actividad de curso actualizada	51824
Usuario ha salido	49837
Envío creado	48695

Fuente: Elaboración de los autores.

De la lista colocada en la tabla 2, se han elegido aquellos eventos que se consideran más relevantes para estimar el rendimiento de los estudiantes. Además, considerando un análisis inicial y general de la información, que se puede obtener de la bitácora del LMS, se obtuvo el total de eventos por año, que se muestran en la figura 2.

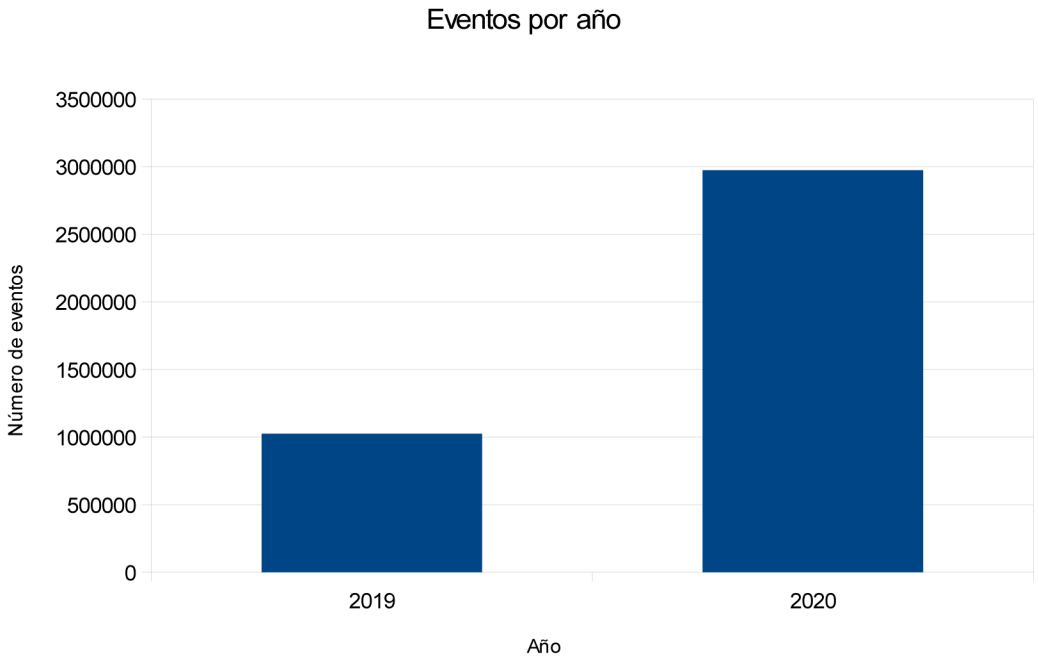

Figura 2. Eventos de la bitácora del LMS por año.
Fuente: Elaboración de los autores.

En la figura 2 se aprecia un aumento importante del uso del LMS institucional. En la figura 3 se muestran los eventos mensuales.

La figura 3 confirma la tendencia al uso cada vez mayor del LMS en los cursos. Nótese el incremento del uso del LMS que coincide con el inicio de la contingencia sanitaria por COVID. Sin embargo, para establecer una relación causal, se requiere

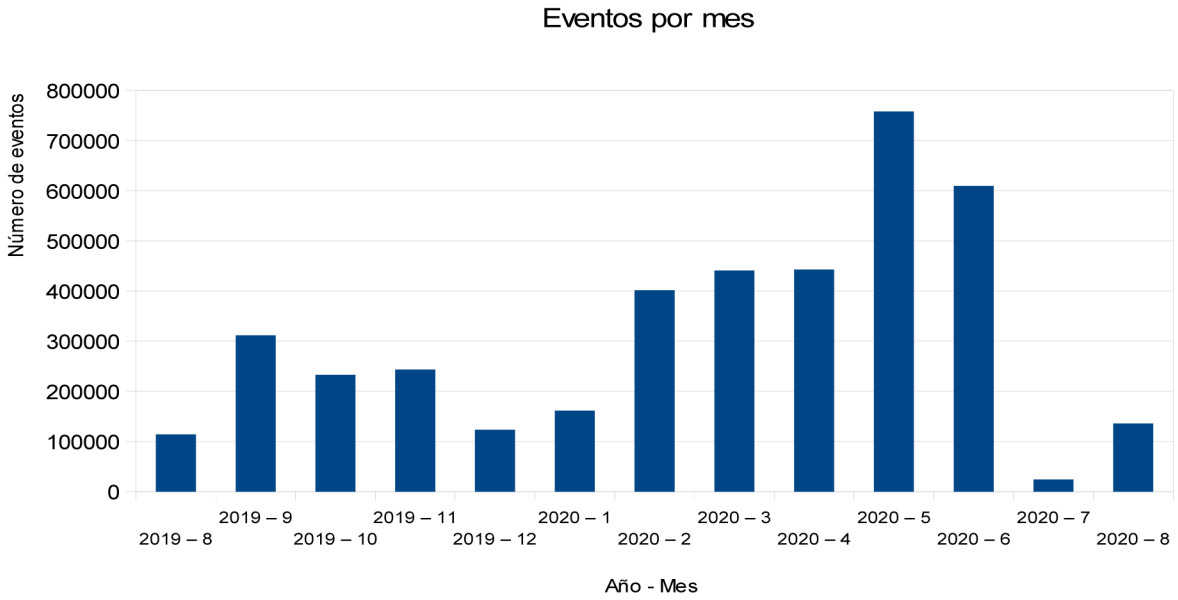

Figura 3. Eventos de la bitácora del LMS por mes.
Fuente: Elaboración de los autores.

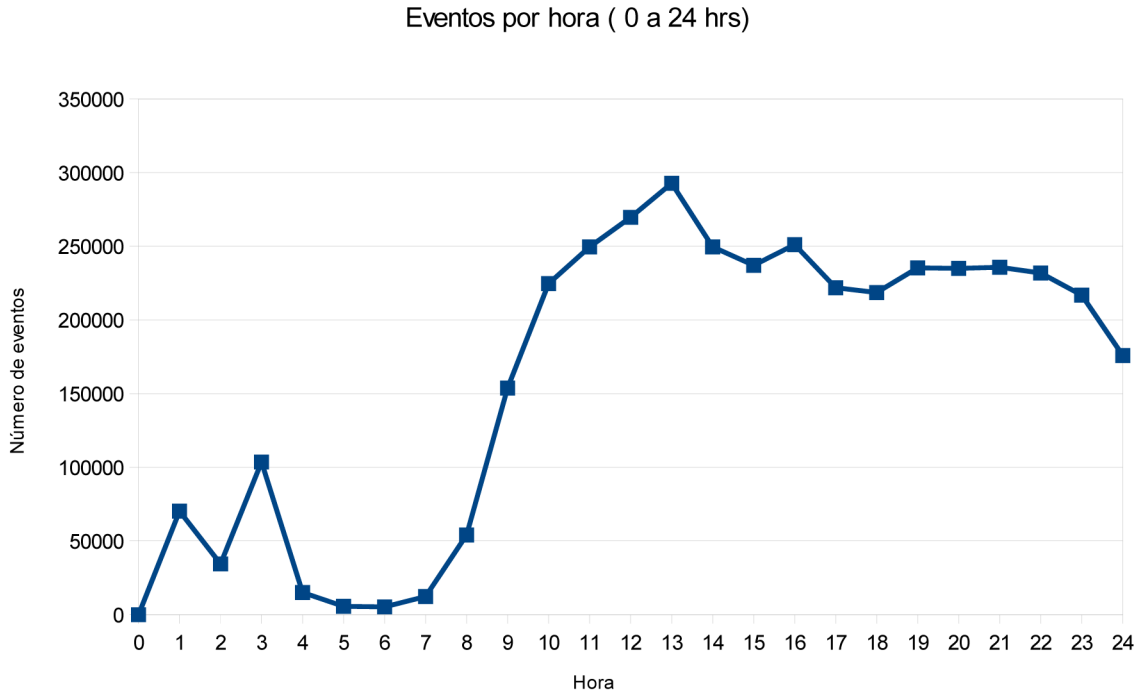

Figura 4. Eventos de la bitácora del LMS por hora.
Fuente: Elaboración de los autores.

Revista Electrónica Científica de Investigación Educativa
Volumen 5, núm. 2, enero-diciembre 2021, pp. 313-333

tener información de más semestres, matrícula estudiantil, aceptación del LMS de parte de los profesores y otros posibles factores. En la figura 4 se muestran los eventos por hora, los cuales permiten identificar las horas de uso más intenso del LMS.

Preparación de los datos

Con el propósito de preparar los datos, los cuales permitan obtener modelos para la predicción del rendimiento de los estudiantes, se identifican eventos que se estiman significativos. Los eventos seleccionados fueron:

- Curso visto.
- El usuario ha guardado un envío.
- Envío creado.
- Intento de examen enviado.
- Suma de eventos totales del estudiante en un curso.

También se obtuvo la calificación final de cada estudiante por curso, información que no se encuentra en la bitácora del LMS. Para ello se consideró una materia, "Programación web", durante los dos semestres del año 2020. Esta última fue cursada por un total de 56 estudiantes.

Para facilitar su preparación e interpretación, las bitácoras en formato CSV se importaron a una base de datos relacional en MySQL. Además, se crearon consultas en la base de datos para obtener la información mostrada en la tabla 3.

Tabla 3. Datos preparados por estudiante y materia.

usuario	accesostotales	envioguardado	enviocreado	intentoexamenenviado	cursovisto	calificacion
1	432	0	13	4	99	95.32
2	332	0	11	1	77	97
3	86	0	0	1	21	0
4	524	3	13	4	124	97.22
5	567	1	12	4	96	84.54
6	459	0	11	2	118	89
7	327	1	8	2	69	50
8	449	1	12	4	116	91.94
9	394	0	12	4	72	91.94
10	613	2	13	4	147	77.85
11	408	0	12	5	103	84.69
12	264	0	4	2	70	50

Fuente: Elaboración de los autores (se omiten los nombres reales de los estudiantes).

Modelado

En esta etapa se utilizan técnicas de ciencia de datos para determinar aquellas que permitan modelar de mejor manera los datos disponibles. Dentro de cada técnica se probarán alternativas de configuración. Algunos de los métodos a tomar en cuenta se mencionan a continuación:

- Técnicas de visualización. Sirven para presentar un conjunto de datos en la forma de gráficos o imágenes, toda vez que ayudan a interpretar la información.
- Regresión simple y múltiple. Se utilizan para modelar las relaciones entre las variables independientes (accesos a la plataforma, entrega oportuna de actividades, etc.) y la variable dependiente (calificación obtenida o esperada).
- Clústeres o agrupamiento. Permite agrupar un conjunto de elementos de datos, tales que los elementos que se encuentran en el mismo grupo (clúster) tienen características similares.

Como ambiente de desarrollo en esta etapa se usa Anaconda (Python, y el editor Spyder).

En primer lugar, se aplican técnicas de visualización. En la tabla 4 se muestran las estadísticas básicas.

Tabla 4. Estadísticas básicas de los datos.

	Accesos totales	Envío guardado	Envío creado	Intento examen enviado	Curso visto	Calificación
count	56.00	56.00	56.00	56.00	56.00	56.00
mean	415.05	.66	8.93	2.66	99.18	67.48
std	185.55	.92	4.39	1.07	59.55	34.51
min	86.00	.00	.00	1.00	21.00	.00
25%	272.25	.00	5.00	2.00	61.00	50.00
50%	420.00	.00	11.00	2.00	95.00	84.62
75%	527.50	1.00	12.00	4.00	122.50	91.99
max	838.00	4.00	13.00	5.00	313.00	99.00

Fuente: Elaboración de los autores (las palabras "accesos" y "eventos" se están usando como sinónimo).

Enseguida se obtuvieron gráficas de frecuencia de los eventos totales, los cuales se muestran en la figura 5.

La figura 6 muestra la distribución del evento "Envío creado", en tanto la figura 7 la distribución del evento "Curso visto".

Revista Electrónica Científica de Investigación Educativa
Volumen 5, núm. 2, enero-diciembre 2021, pp. 313-333

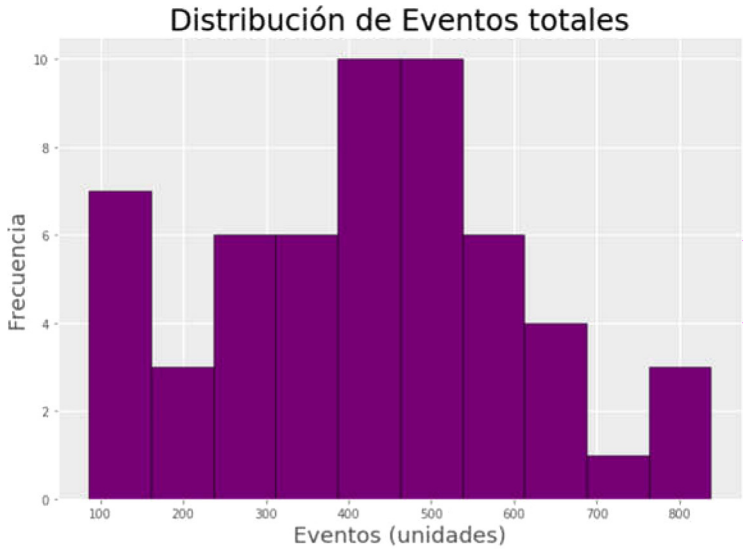

Figura 5. Gráfica de frecuencia de eventos totales.
Fuente: Elaboración de los autores.

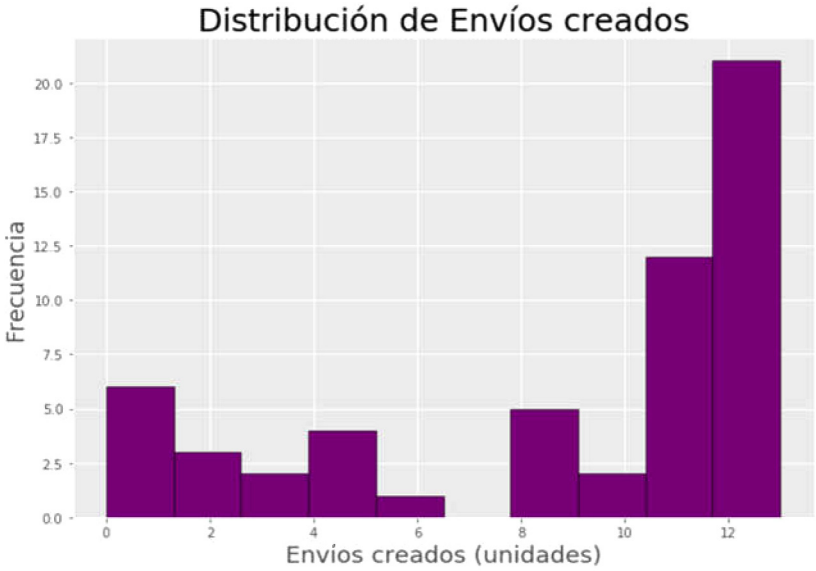

Figura 6. Gráfica de frecuencia de "Envío creado".
Fuente: Elaboración de los autores.

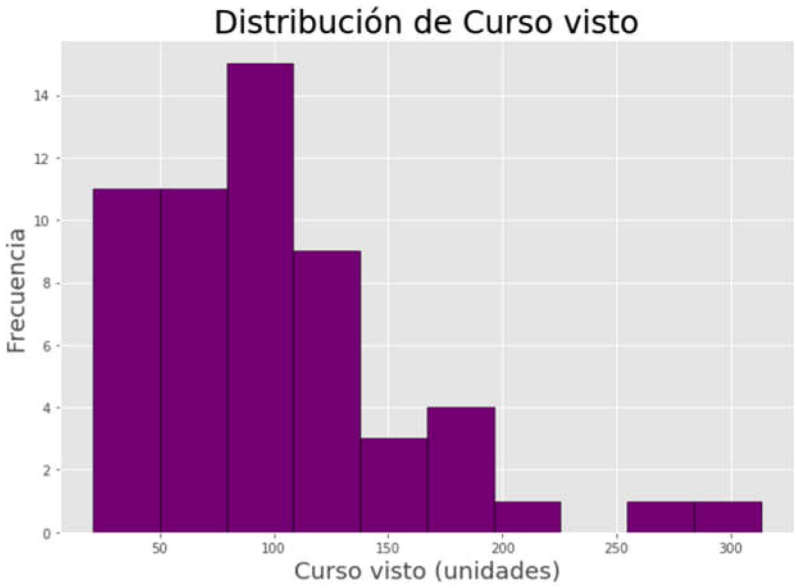

Figura 7. Gráfica de frecuencia de "Curso visto".
Fuente: Elaboración de los autores.

Se procede enseguida a la determinación de regresiones simples de las variables independientes (diferentes tipos de eventos, de uno en uno) contra la calificación final obtenida por los estudiantes (variable dependiente).

En la primera regresión se ha considerado el campo calculado "Eventos totales" contra la calificación. Los resultados se aprecian en la gráfica de la figura 8.

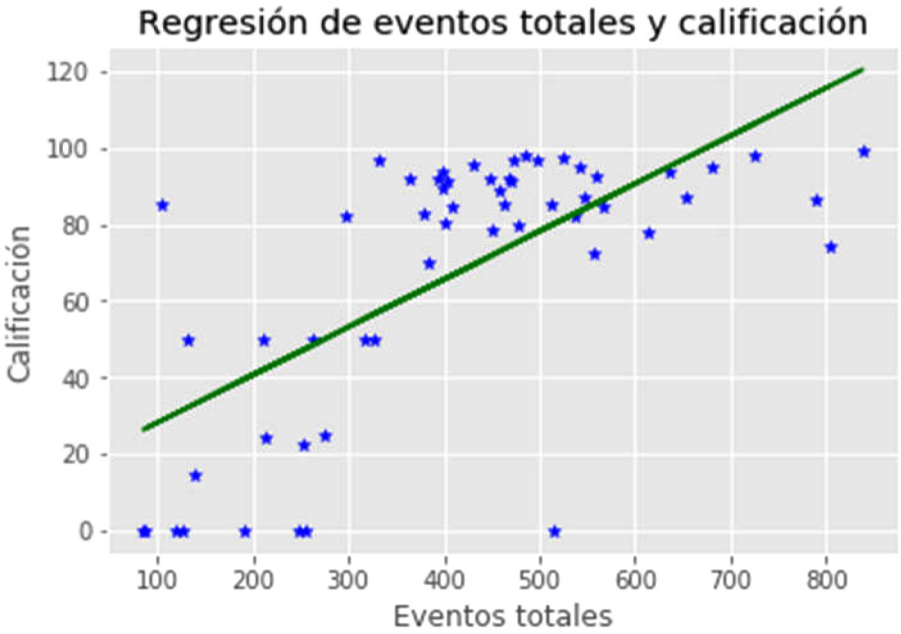

Figura 8. Gráfica de regresión de "Eventos totales" y la "Calificación".
Fuente: Elaboración de los autores.

Revista Electrónica Científica de Investigación Educativa
Volumen 5, núm. 2, enero-diciembre 2021, pp. 313-333

Los indicadores de esta última regresión se muestran enseguida:

- coefficient of determination: 0.45196380794014623
- intercept: 15.581849106360984
- slope: [0.1250461]

Obsérvese que la variable "Eventos totales" tiene un coeficiente de determinación de 0.45, que es el segundo valor mayor de todas las regresiones simples. Este coeficiente indica qué tan bien los resultados observados son replicados por el modelo.

En la figura 9 se puede ver la gráfica de regresión de "Envío guardado" y "Calificación".

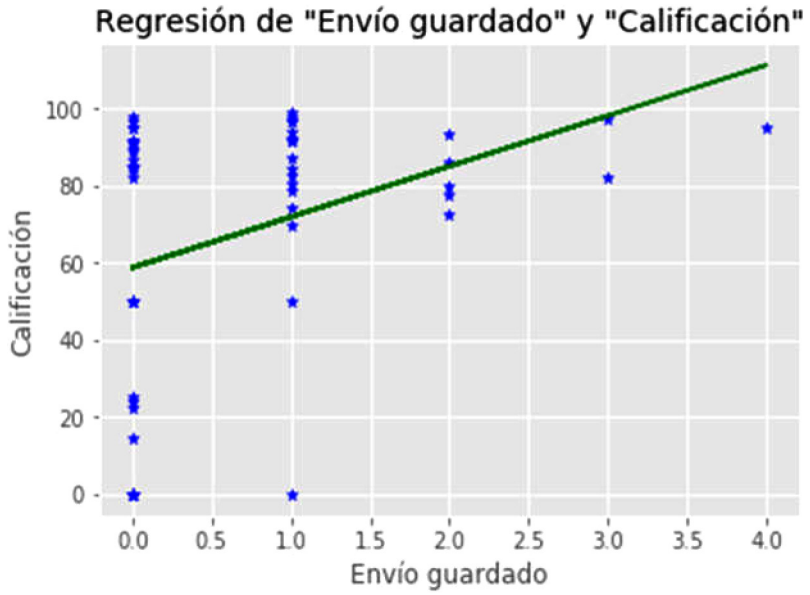

Figura 9. Gráfica de regresión de "Envío guardado" y la "Calificación".
Fuente: Elaboración de los autores.

Los indicadores de esta regresión se muestran enseguida:

- coefficient of determination: 0.12255227577197214
- intercept: 58.80604526275413
- slope: [13.13220176]

El valor del coeficiente de determinación indica que el atributo "Envío guardado" es de poco peso para estimar la calificación.

En la figura 10 se muestra la gráfica de regresión de "Envío creado" y "Calificación".

Los indicadores de esta regresión se muestran enseguida:

- coefficient of determination: 0.6713403666951503
- intercept: 9.963546778107293
- slope: [6.44214276]

El atributo "Envío creado" presenta el valor mayor en su coeficiente de determinación, respecto a la calificación; por lo tanto, en este caso es el atributo de mayor peso para determianr la calificación.

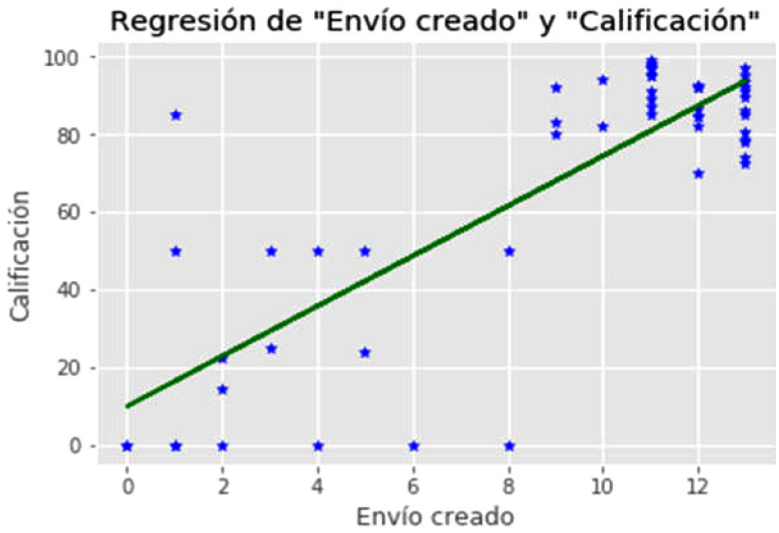

Figura 10. Gráfica de regresión de "Envío creado" y la "Calificación".
Fuente: Elaboración de los autores.

En la figura 11 se muestra la gráfica de regresión de "Curso visto" y "Calificación".

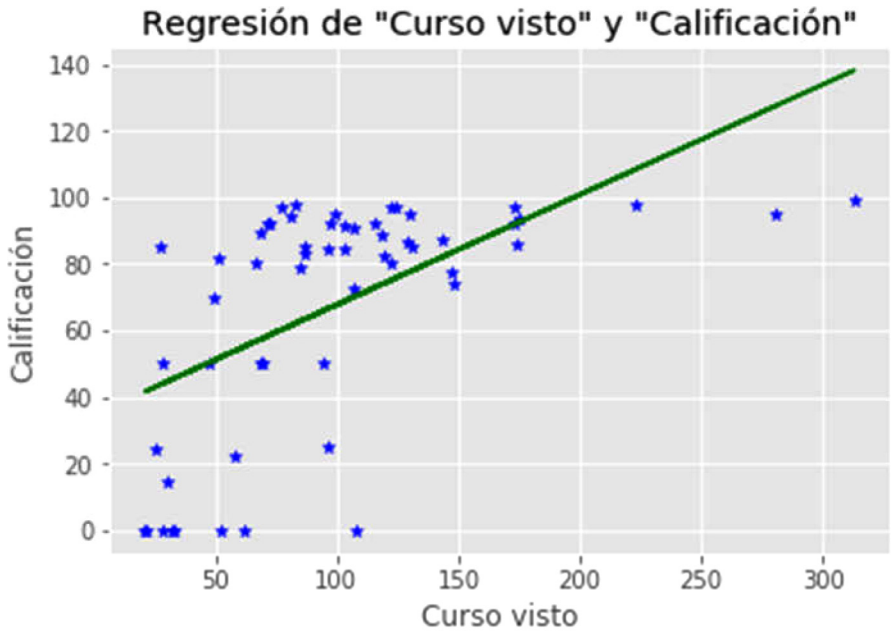

Figura 11. Gráfica de regresión de "Curso visto" y la "Calificación".
Fuente: Elaboración de los autores.

Los indicadores de esta regresión se muestran enseguida:

Revista Electrónica Científica de Investigación Educativa
Volumen 5, núm. 2, enero-diciembre 2021, pp. 313-333

- coefficient of determination: 0.32521554456224044
- intercept: 34.70442085748029
- slope: [0.33049738]

Con base en los resultados de las regresiones simples, el orden de los atributos, del mayor al menor para la determinación de la calificación es: "Envío creado", "Accesos (eventos) totales", "Curso visto" y "Envío guardado".

Se procedió a realizar un análisis de regresión múltiple considerando las variables independientes "Accesos totales", "Envío creado" y "Curso visto". Se obtuvo la siguiente expresión para la regresión múltiple:

Calificación = 10.946203179485167 +
("Accesos totales" * -0.0640666675991691) + ("Envío creado" * 6.90473320908922)
+ ("Curso visto" * 0.21656060000210814)

Con los siguientes estadísticos:

- coefficient of determination: 0.701085772015392
- intercept: 10.946203179485167
- slope: [-0.06406667 6.90473321 0.2165606]

Para la regresión múltiple se probaron diferentes combinaciones de atributos, obteniéndose, en todos los casos, un coeficiente de determinación entre 0.70 y 0.71, lo que se considera aceptable para realizar una predicción.

Clústeres o agrupamiento

En la siguiente etapa se analizaron los datos a través de un algoritmo no supervisado, en este caso el de agrupamiento, utilizando el método k-means.

Con este método se desea verificar qué tan bien se identifican tres grupos de estudiantes (regulares, riesgo medio y riesgo alto, en términos académicos), además de que los tres grupos representen con la mayor exactitud posible la situación real de cada estudiante. Al ser un método no supervisado, los datos de entrada serán únicamente las variables independientes, es decir, no se incluye la calificación real.

Después de aplicar el método se obtuvo la gráfica de la figura 12, en la cual se observa que se solicita determinar tres grupos.

Para la obtención de los grupos se siguieron las recomendaciones para el uso del k-means. Uno de ellos es la normalización o estandarización de los datos, es decir, procurar que los valores de las variables se encuentren en el rango de 0 y 1. Los grupos deseados se proporcionaron directamente al método, usando también el algoritmo para determinar la cantidad óptima de grupos, mediante el "método del codo". Para obtener la gráfica se aplicó el método de componentes principales (PCA), el cual consiste en una simplificación de las variables, así como determinar cuáles tienen un mayor peso en la identificación de los grupos. En la gráfica se reconocen como "componente 1" y "componente 2". También se obtuvieron, y muestran, los centros de cada grupo, identificados con la letra "X".

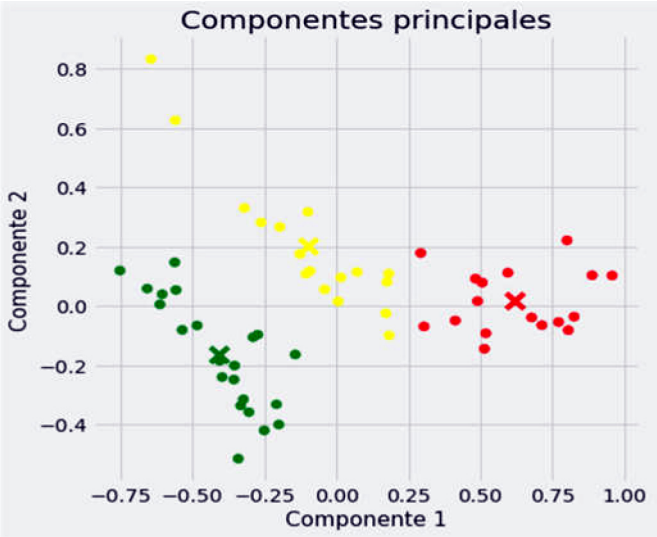

Figura 12. Clústeres obtenidos con el método k-means.
Fuente: Elaboración de los autores.

El método k-means es no-determinístico, ello implica que, en cada ejecución, para los mismos datos de entrada, se pueden obtener diferentes asignaciones de clúster. En los parámetros del método se eligen siempre tres clústeres, lo cual se observó en las diferentes ejecuciones. Ello se debe a que en el método varía la asignación del número de clúster para cada grupo, mas siempre asignó los mismos estudiantes dentro de cada grupo.

Para identificar el clúster en el cual fue clasificado cada estudiante, se obtuvo el archivo de datos generado por el método PCA. En este se indica el número de registro, coordenadas del componente 1, del componente 2 y el número de clúster asignado, tal como se muestra en la tabla 5 (se distinguen solo los primeros 12 renglones).

Tabla 5. Resultado tabular del método PCA.

	Componente_1	Componente_2	KMeans_Clusters
0	-0.3033564442	-0.3591797048	0
1	0.1819910959	0.1082342613	2
2	0.9575972216	0.1018853242	1
3	-0.6042375732	0.0388730481	0
4	-0.4052598061	-0.1856883864	0
5	-0.0403366691	0.0549555858	2
6	0.1757528334	0.0803109004	2
7	-0.353393441	-0.2016714979	0
8	-0.1994819999	-0.4009319242	0
9	-0.6122202904	0.0044831503	0
10	-0.3406945611	-0.5158782313	0
11	0.490500523	0.0156793484	1

Fuente: Elaboración de los autores.

Revista Electrónica Científica de Investigación Educativa
Volumen 5, núm. 2, enero-diciembre 2021, pp. 313-333

Se obtuvieron estadísticas de apoyo para la interpretación de los clústeres. Además, se comparan los clústeres obtenidos contra los datos de los estudiantes, logrando la siguiente información, que se aprecia en las tablas 6a, 6b y 6c.

Tabla 6a. Estadísticas del clúster 0 (color verde en la figura 12).

Análisis de clústeres		
Clúster 0:	Alumnos	22
	Promedio accesos totales	516.64
	Promedio envío creado	12.55
	Promedio intento examen enviado	3.86
	Promedio curso visto	112.27
	Promedio calificaciones	85.63

Fuente: Elaboración de los autores.

Tabla 6b. Estadísticas del clúster 1 (color rojo en la figura 12).

Clúster 1:	Alumnos	17
	Promedio accesos totales	208.24
	Promedio envío creado	2.82
	Promedio intento examen enviado	1.82
	Promedio curso visto	59.32
	Promedio calificaciones	21.84

Fuente: Elaboración de los autores.

Tabla 6c. Estadísticas del clúster 2 (color amarillo en la figura 12).

Clúster 2:	Alumnos	17
	Promedio accesos totales	490.41
	Promedio envío creado	10.35
	Promedio intento examen enviado	1.94
	Promedio curso visto	132.41
	Promedio calificaciones	89.65

Fuente: Elaboración de los autores.

- Clúster 0 (color verde). Se observa que es el clúster con un promedio mayor en la variable "Envío creado", que obtuvo el coeficiente de determinación mayor de las variables analizadas y por lo tanto la de mayor importancia para predecir una calificación final. En la gráfica de "Componentes principales" todos los elementos se encuentran cercanos. Todos los alumnos del clúster aprobaron el curso, por lo que este clúster se puede considerar como de alumnos "regulares" o bien significa que tienen mayores posibilidades de éxito.

- Clúster 1 (color rojo). Se observa que es aquel que cuenta con el menor promedio en la variable "Accesos totales", además de promedios menores en todas las demás variables analizadas. En la gráfica de "Componentes principales" también sus elementos se encuentran cercanos. De los 17 alumnos del clúster, 16 de ellos no alcanzaron calificación aprobatoria, por lo que este clúster se puede considerar como de alumnos de "alto riesgo" o que tienen mayores posibilidades de fracaso. Son alumnos que deben ser identificados a tiempo para una posible intervención académica de apoyo.
- Clúster 2 (color amarillo). La gráfica de "Componentes principales" tiene una dispersión mayor (por dos estudiantes que se encuentran relativamente lejos del centro del clúster). Tiene características interesantes, por un lado, su promedio final de calificaciones fue el mayor de los tres grupos. También el promedio de la variable "Curso visto" fue el mayor. Pero en la variable "Envío creado" su promedio fue inferior al del clúster 0, recuérdese que dicha variable mostró el mayor coeficiente de determinación. De los 17 alumnos del total de este grupo, 16 de ellos lograron aprobar el curso y solo uno de ellos no aprobó. Por lo anterior, este clúster se puede considerar de "riesgo medio".

Evaluación

De los resultados obtenidos, se considera que el método de regresión múltiple, de lado del método de agrupamiento k-means, sirve para predecir el rendimiento de los estudiantes. En esta dirección, se considera que el coeficiente de determinación de 0.70 es aceptable. Por otra parte, el método de clúster permitió identificar en el clúster 1 a los estudiantes con alto grado de riesgo académico (de 17 estudiantes en este grupo, 16 no acreditaron el curso).

En consecuencia, consideramos necesaria más información para mejorar los modelos, por ejemplo, tener disponible la base de datos del LMS, en la cual existe mayor información, como calificaciones de actividades concretas, fechas de entrega programadas contra las reales, participaciones en foros, chats y otras actividades asignadas por los profesores.

Despliegue

Para que los modelos obtenidos sean de utilidad, es necesario que se apliquen y desplieguen en la institución. Además es necesaria una política institucional de gestión del *Learning Analytics*, considerando todos sus aspectos.

Un elemento del despliegue lo constituyen las aplicaciones informáticas que se utilizarán de lado de los modelos obtenidos, las cuales permitirán que los usuarios designados por la institución las utilicen de una manera sencilla.

Como ejemplo de una de las interfaces, en las figuras 13a y 13b se muestra parte de la aplicación web con la lista de estudiantes del curso "Programación web" y un

Revista Electrónica Científica de Investigación Educativa
Volumen 5, núm. 2, enero-diciembre 2021, pp. 313-333

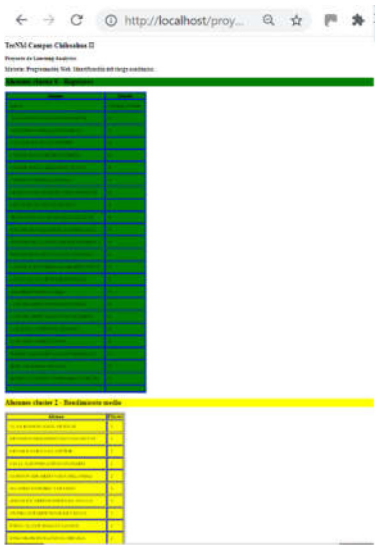

Figura 13a. Aplicación web mostrando clústeres.
Fuente: Elaboración de los autores.

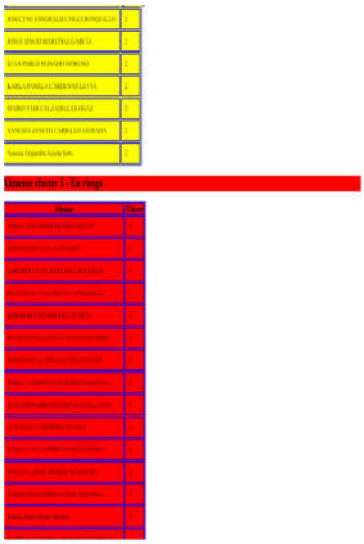

Figura 13b. Aplicación web mostrando clústeres.
Fuente: Elaboración de los autores.

agrupamiento y color asignado según el clúster al que pertenecen. A propósito, no se puede leer el nombre real de los alumnos por motivo de privacidad, que un profesor o tutor autorizado sí puede ver. Se resalta en esta figura el uso de los colores para indicar el nivel de desempeño esperado de los estudiantes, en verde los "regulares", en amarillo los de "riesgo medio" y en rojo los alumnos en "riesgo alto".

En la figura 14 se muestra la parte de la aplicación que usa el modelo de regresión para estimar la calificación de un estudiante.

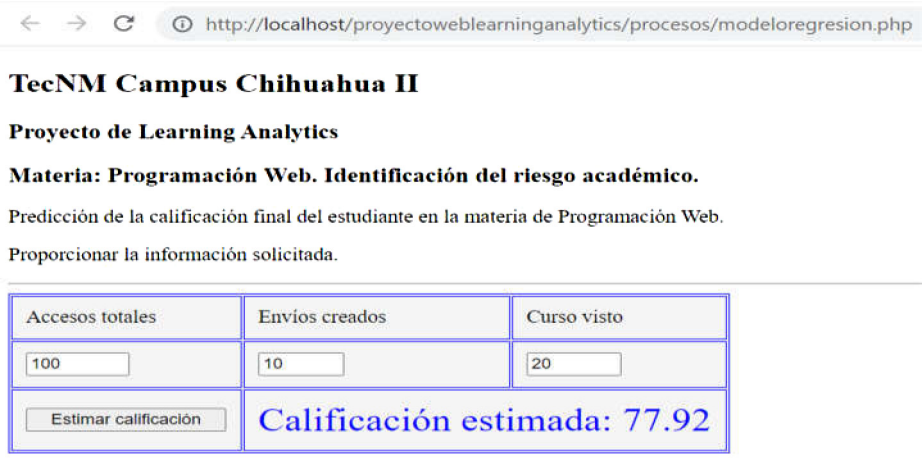

Figura 14. Aplicación web que usa la regresión.
Fuente: Elaboración de los autores.

Mediante la aplicación, el tutor o personal designado por la institución puede identificar a los estudiantes con riesgo académico alto y puede servirle de apoyo en la decisión de una intervención académica en favor de ellos.

Conclusiones y recomendaciones

El avance actual del proyecto demuestra su factibilidad técnica. Mediante técnicas de LA es posible predecir con cierto grado de certeza el rendimiento de un estudiante, así como identificar a aquellos que tienen alto riesgo académico, de modo que esta información pueda servir de apoyo para decidir sobre una intervención en un momento oportuno.

Con las estadísticas básicas obtenidas de las bitácoras del LMS, aunadas al método k-means, es posible obtener información útil. Por ejemplo, dado un estudiante, conocer el clúster al que pertenece, comparar su información contra los demás estudiantes del mismo clúster, así como comparar sus datos contra otros clústeres y el grupo total. Si esta información se hace visible al estudiante, le podría servir de incentivo para mejorar sus indicadores. Esta misma información sería útil también para profesores y tutores.

Respecto a las limitantes del trabajo, este se aplicó solamente en una materia, durante dos semestres y con el mismo profesor; reconociendo que cada materia o profesor tiene características particulares que deben ser analizadas y consideradas para obtener un modelo más robusto y genérico, o bien que este se adapte a cualquier otro caso.

Respecto a la fuente de información, se utilizaron solo las bitácoras del LMS y la calificación final obtenida en el curso. Para obtener un mejor modelo deben considerarse otras fuentes de datos tales como:

- Base de datos del LMS.
- Sistema de información escolar.
- Uso de bibliotecas digitales escolares.

La base de datos del LMS puede proporcionar información como las fechas programadas de entrega de actividades contra las fechas reales de entrega, las calificaciones asignadas a las actividades particulares, resultados de exámenes, participaciones en foros, chats y otras actividades.

El sistema de información escolar puede proporcionar información tal como el historial de calificaciones, escuela de procedencia, promedio del nivel anterior, datos económicos y sociales. En cuanto al uso de las bibliotecas digitales, pudieran ser de interés las fuentes consultadas, las acciones como copiar un texto, imprimir, tiempo de lectura, entre otras.

Otra limitante es que el modelo se debe ajustar para ser capaz de predecir el rendimiento en las etapas iniciales y medianas de los cursos, cuando aún no se tiene la información completa. Los modelos obtenidos en este trabajo usan la información completa del semestre, incluida la calificación final obtenida. Los métodos de regresión se podrían aplicar solamente en caso de poder obtener calificaciones parciales, información no disponible en la bitácora del LMS, solamente se encuentra en la base de datos. El método k-means puede aplicarse con datos parciales del semestre, para

ello es necesario realizar pruebas en diferentes etapas y comparar resultados contra el método aplicado con toda la información.

Todo proyecto de *Learning Analytics* debe considerar el aspecto ético, en particular la privacidad de la información, esto dentro de los acuerdos y políticas institucionales sobre el uso de los datos. Este es un aspecto aún no abordado en el presente trabajo.

Para que un proyecto de este tipo tenga éxito, debe realizarse en forma institucional y considerando a las personas involucradas en todas las etapas del mismo. Se debe trabajar en demostrar las bondades del LA a directivos, docentes y estudiantes, de manera que el proyecto no sea solo de un grupo limitado de personas, sino que se pueda asumir como un proyecto de toda la institución. Solo de esta manera se puede esperar tener un sistema LA que sea de real beneficio para las instituciones.

Referencias

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., y Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. En *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 415-421). Recuperado de: https://dl.acm.org/doi/abs/10.1145/3041021.3054164.

Dawson, S., Joksimovic, S., Poquet, O., y Siemens, G. (2019). Increasing the impact of learning analytics. En *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 446-455). Recuperado de: https://www.researchgate.net/profile/Srecko_Joksimovic/publication/331333276_Increasing_the_Impact_of_Learning_Analytics/links/5c870c7aa6fdcc88c39bf70e/Increasing-the-Impact-of-Learning-Analytics.pdf.

Gasevic, D., Tsai, Y. S., Dawson, S., y Pardo, A. (2019). How do we start? An approach to learning analytics adoption in higher education. *The International Journal of Information and Learning Technology*, *36*(4), 342-353. Recuperado de: https://research.monash.edu/files/296208063/277681827_oa.pdf.

Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z., y Hlosta, M. (2019). A large-scale implementation of predictive learning analytics in higher education: The teachers' role and perspective. *Educational Technology Research and Development*, *67*(5), 1273-1306. Recuperado de: https://link.springer.com/article/10.1007/s11423-019-09685-0.

Hwang, G.-J., Chu, H. C., y Yin, C. (2017). Objectives, methodologies and research issues of learning analytics. *Interactive Learning Environments*, *25*(2). Recuperado de: https://www.tandfonline.com/doi/full/10.1080/10494820.2017.1287338.

Martínez Navarro, A., y Moreno-Ger, P. (2018). Comparison of clustering algorithms for learning analytics with educational datasets. *IJIMAI*, *5*(2), 9-16. Recuperado de: https://dialnet.unirioja.es/descarga/articulo/6907743.pdf.

Romero, C., y Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1355. Recuperado de: https://bookdown.org/chen/la-manual/files/Romero%20and%20Ventura%20-%202020.pdf.

Selwyn, N. (2019). What's the problem with Learning Analytics? *Journal of Learning Analytics*, *6*(3), 11-19. Recuperado de: https://learning-analytics.info/index.php/JLA/article/download/6386/7308.

Sclater, N., Peasgood, A., y Mullan, J. (2016). Learning analytics in higher education. Londres: Jisc. Recuperado de: https://www.academia.edu/download/61301572/learning-analytics-in-he-v320191122-83155-kaxmyg.pdf.

Simanca, F. A., Herrera, R., Crespo, L. G., Baena, R., y Burgos, D. (2019). A solution to manage the full life cycle of learning analytics in a learning management system: AnalyTIC. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje*, *14*(4), 127-134. Recuperado de: https://ieeexplore.ieee.org/abstract/document/8894865.

Viberg, O., Hatakka, M., Bälter, O., y Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98-110. Recuperado de: https://www.sciencedirect.com/science/article/pii/S0747563218303492.

Wikipedia (2022, may. 9). *Analítica de aprendizaje*. Recuperado de: https://es.wikipedia.org/w/index.php?title=Anal%C3%ADtica_de_aprendizaje&oldid=137026001.

Wirth, R., y Hipp, J. (2000, abr.). CRISP-DM: Towards a standard process model for data mining. En *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (vol. 1). Londres: Springer-Verlag. Recuperado de: www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

Cómo citar este artículo:

Nevárez Chávez, L., Caldera Franco, M. I., y Ronquillo Máynez, G. (2021). Obtención de un modelo de *Learning Analytics* con información de un LMS. *RECIE. Revista Electrónica Científica de Investigación Educativa*, 5(2), pp. 313-333. doi: doi.org/10.33010/recie.v5i2.1314.