Tipo de artículo: Artículo original

Predicción del rendimiento académico de estudiantes universitarios en la plataforma Moodle mediante Big Data

Predicting the academic performance of university students on the Moodle platform using Big Data

Yandry José Olarte Sancán[1*], https://orcid.org/0000-0002-9581-5557
Marely del Rosario Cruz Felipe[2], https://orcid.org/0000-0003-1937-1568
Jenmer Maricela Pinargote Ortega [3], https://orcid.org/0000-0002-4018-9616
Juan Andrés Mejía Almenaba[4], https://orcid.org/0000-0002-2301-4750

[1] Facultad de Ciencias Básicas, Universidad Técnica de Manabí. yandry.olarte@utm.edu.ec
[2] Facultad de Ciencias Informáticas, Universidad Técnica de Manabí. marely.cruz@utm.edu.ec
[3] Facultad de Ciencias Informáticas, Universidad Técnica de Manabí. maricela.pinargote@utm.edu.ec
[4] Facultad de Ciencias Administrativas y Económicas, Universidad Técnica de Manabí. juan.mejia@utm.edu.ec

*** Autor para correspondencia: yandry.olarte@utm.edu.ec**

Resumen

El objetivo del presente trabajo de investigación fue analizar el rendimiento académico de los estudiantes universitario de pregrado en la plataforma Moodle utilizando Big Data, con el fin de desarrollar un modelo predicción. La metodología seleccionada comprende tres etapas fundamentales: el análisis, extracción y almacenamiento de datos, definición de variables y filtrado de datos y la fase de predicción para obtener un modelo predictivo del rendimiento académico. Se utilizó herramientas Big Data y el lenguaje Python junto con algoritmos supervisados como XGBoost, CatBoost, Random Forest, Decision Tree, SVC, K-vecino, Naive Bayes, y Logistic Regression para optimizar el modelo. La calidad de predicción del modelo se evaluó mediante métricas de precisión, sensibilidad y exactitud. Los resultados revelaron que los algoritmos Random Forest, CatBoost y XGBoost presentaron los mejores rendimientos en la predicción, con una precisión del 94%, 93% y 92%, respectivamente, tanto en el entrenamiento como en las pruebas. Aunque se seleccionó el algoritmo Random Forest, el análisis no paramétrico de Wilcoxon demostró que los tres algoritmos presentaban un rendimiento similar, lo que indica que cualquiera de ellos podría ser utilizado sin diferencias significativas. Estos resultados demuestran que, a partir del análisis de los datos de las interacciones de los estudiantes con la plataforma, es posible obtener información y conocimiento valioso y oportuno para predecir su rendimiento académico. Esta información puede resultar relevante en la toma de decisiones para mejorar el proceso de enseñanza-aprendizaje en entornos virtuales.

Palabras clave: Rendimiento académico; Big Data; modelo predictivo; Moodle, minería de datos.

Abstract

The objective of this research work was to analyze the academic performance of undergraduate university students in the Moodle platform using Big Data, in order to develop a predictive model. The selected methodology comprises three fundamental stages: analysis, extraction and storage of data, definition of variables and data filtering, and the prediction phase to obtain a predictive model of academic performance. Big Data tools and the Python language were used along with supervised algorithms such as XGBoost, CatBoost, Random Forest, Decision Tree, SVC, K-neighbor, Naive Bayes, and Logistic Regression to optimize the model. The prediction quality of the model was evaluated using precision, sensitivity, and accuracy metrics. The results revealed that the Random Forest, CatBoost, and XGBoost algorithms presented the best prediction performances, with 94%, 93%, and 92% accuracy, respectively, in both training and testing. Although the Random Forest algorithm was selected, the Wilcoxon

nonparametric analysis showed that all three algorithms performed similarly, indicating that any of them could be used without significant differences. These results demonstrate that, from the analysis of the data of students' interactions with the platform, it is possible to obtain valuable and timely information and knowledge to predict their academic performance. This information can be relevant in decision making to improve the teaching-learning process in virtual environments.

Keywords: *Academic performance; Big Data; predictive modeling; Moodle, data mining.*

Recibido: 18/05/2023
Aceptado: 28/08/2023
En línea: 01/09/2023

Introducción

La integración de la tecnología en la educación ha transformado la forma en que los estudiantes acceden al conocimiento y los docentes gestionan el proceso de enseñanza-aprendizaje. En este contexto, las plataformas de aprendizaje en línea, como Moodle, han ganado popularidad en instituciones educativas de todo el mundo (Gámez et al., 2018). Según (Belmonte et al., 2019), Moodle se ha convertido en una de las plataformas más utilizadas en el ámbito educativo, proporcionando un entorno virtual de aprendizaje que facilita la colaboración, la comunicación y la evaluación de los estudiantes. Estas plataformas brindan un entorno virtual que facilita la interacción entre estudiantes y docentes, el acceso a materiales de estudio y la evaluación del rendimiento académico (Muñoz-Gea et al., 2016). Sin embargo, a pesar de sus beneficios, la predicción del rendimiento académico de los estudiantes en estas plataformas sigue siendo un desafío (Menacho Chiok, 2017).

El análisis de grandes volúmenes de datos, conocido como Big Data, ha surgido como una herramienta poderosa para comprender y predecir diversos fenómenos. En el ámbito educativo, el uso de Big Data ha mostrado prometedores resultados en la predicción del rendimiento académico de los estudiantes. Según, La aplicación de técnicas de Big Data en la educación puede ayudar a los docentes a identificar patrones ocultos en los datos y tomar decisiones informadas para mejorar el rendimiento académico de los estudiantes (Saiz Manzanares et al., 2018). La plataforma Moodle, al recopilar una gran cantidad de datos sobre las interacciones de los estudiantes, se convierte en un recurso valioso para aplicar técnicas de Big Data y mejorar la toma de decisiones en el ámbito educativo (Menacho, 2020).

En este artículo, se pretende abordar el tema de la predicción del rendimiento académico de los estudiantes universitarios en la plataforma Moodle mediante el uso de Big Data. Para ello, se realizó una revisión de la literatura académica especializada, que incluye estudios e investigaciones relevantes en este campo. Además, se analizaron las diferentes técnicas y enfoques utilizados en la predicción del rendimiento académico, así como las variables consideradas en estos modelos predictivos (Miranda & Guzmán, 2017).

Uno de los aspectos fundamentales a considerar en la predicción del rendimiento académico es la selección adecuada de variables predictoras (Yang et al., 2020). Según (Roldan & Castro, 2016), los datos demográficos de los estudiantes, el historial académico previo, la participación en actividades en la plataforma, así como los patrones de acceso a los recursos educativos, son algunas de las variables más utilizadas en la predicción del rendimiento académico en entornos virtuales de aprendizaje lo cual coincide con varios autores (Akçapınar, 2016; Bogarín Vega et al., 2015; Hidalgo Cajo, 2018; E. López et al., 2018; Menacho Chiok, 2017). La combinación de estas variables, junto con algoritmos de aprendizaje automático y técnicas de minería de datos, permiten obtener modelos predictivos que pueden identificar patrones y tendencias en el rendimiento académico.

La predicción del rendimiento académico de los estudiantes en la plataforma Moodle no solo puede ser útil para los docentes, sino también para los propios estudiantes. Según (J. López et al., 2020), la predicción del rendimiento académico puede ayudar a los estudiantes a identificar áreas de mejora y adoptar estrategias de estudio más efectivas, lo que contribuye a un mejor desempeño y éxito académico. Además, los resultados de estas predicciones pueden ser utilizados por las instituciones educativas para implementar intervenciones personalizadas y mejorar la calidad de la educación.

La Universidad Técnica de Manabí (UTM) es una Institución de Educación Superior que ha incursionado en el campo de la virtualidad desde el año 2014, actualmente oferta un total de 52 carreras, 44 presencial y 8 en línea, por lo que para llevar a cabo las actividades educativas ha implementado diversas herramientas tecnológicas.

El uso de herramientas tecnológicos y sistemas en la UTM a partir del año 2020 tuvo un incremento exponencial producto de la pandemia mundial provocada por el COVID-2019, obligando al profesorado de modalidad presencial a adaptar los modelos de enseñanza tradicional a un ambiente virtual, para atender la necesidad de aproximadamente 26118 estudiantes de carreras presenciales.

El cambio a la modalidad virtual ha provocado que se genere grandes volúmenes de datos producto de las interacciones de usuarios con la plataforma virtual, estos datos no son aprovechados, debido a que no existen mecanismos de análisis de datos aplicados a la plataforma que permita conocer el comportamiento de los estudiantes en entornos virtuales y predecir su rendimiento académico.

El rendimiento académico, se ve reflejado en las calificaciones obtenidas por los estudiantes; para el periodo académico 2021, de acuerdo al reporte del total de inscritos/reprobados, proporcionado por el (SGA & UTM, 2021) se evidencia que del total de estudiantes inscritos por asignatura en promedio existe un 16% de estudiantes que reprobó una materia.

El análisis de los registros que almacena Moodle en forma de logs puede ayudar a los profesores y diferentes autoridades a conocer mejor la actividad desarrollada por sus estudiantes en los espacios virtualizados, saber si existen distintos tipos de estudiantes de acuerdo a su comportamiento, y predecir cuáles van asociados a su rendimiento académico.

Por tal razón, es necesario la aplicación de mecanismos de Big Data a la plataforma Moodle; debido a que se carece de procedimientos de análisis del gran volumen de datos, además de examinar los principales algoritmos aplicados en estudios similares y replicarlos para generar un modelo de predicción del rendimiento académico que se adapte a la realidad de la UTM y de esta manera contribuir a la mejora del proceso de enseñanza - aprendizaje en el entorno virtual.

Materiales y métodos

La metodología empleada para el desarrollo del trabajo de investigación toma como base una ya existente como es la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), la cual fue adaptada considerando las necesidades y requerimientos para el manejo de grandes volúmenes de datos, la cual incluye como primera fase el análisis, extracción y almacenamiento de datos, luego en la segunda fase la definición de variables y filtrado de datos, para luego continuar con la última fase, la predicción, tal como se observa en la figura 1.

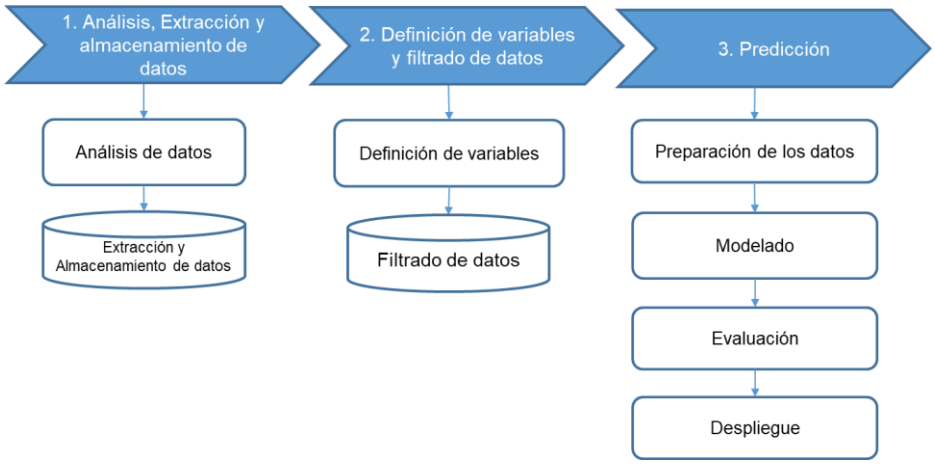

Figura 1. Etapas de la metodología de investigación utilizada.

El desarrollo de esta metodología aporta en el análisis del comportamiento de los estudiantes con la plataforma Moodle y predicción del rendimiento académico que incluye la interacción de diversas herramientas de Big Data como se muestra en la figura 2.

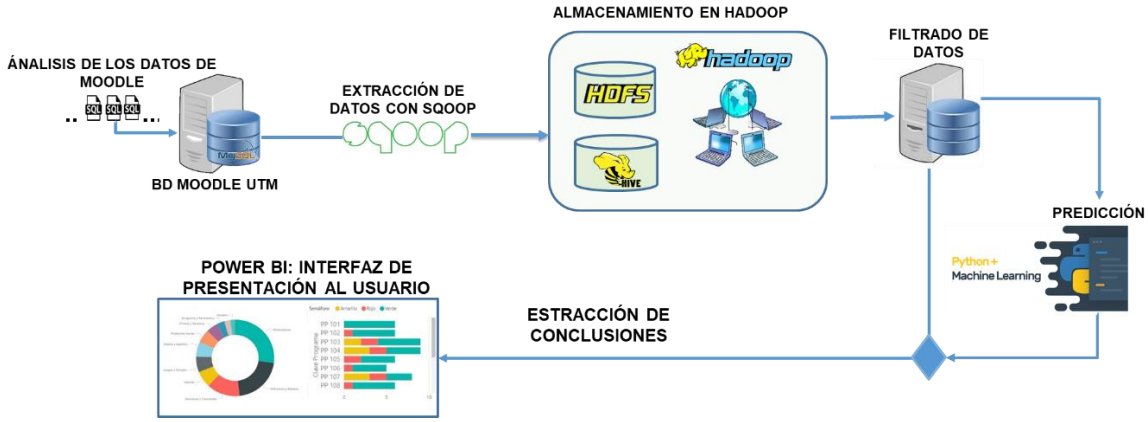

Figura 2. Interacción de varias herramientas de Big Data.

En la primera fase se realizó un análisis exploratorio de la base de datos relacional de Moodle que contiene 462 tablas con la información de los diversos cursos, con la finalidad de comprender el modelo de datos sobre el que trabaja la plataforma y se estableció las tablas relevantes para el estudio, como se muestra en la tabla 1.

Tabla 1. Identificación de tablas de Moodle más relevantes para el estudio.

	Tablas de Moodle
	mdl_logstore_standard_log
Conjunto de tablas I	mdl_user mdl_assign, mdl_assign_submission, mdl_config, mdl_context, mdl_course, mdl_course_modules, mdl_forum, mdl_forum_discussions, mdl_forum_posts, mdl_grade_grades, mdl_grade_items, mdl_groups, mdl_groups_member, mdl_modules, mdl_quiz, mdl_quiz_attempts, mdl_role_assignments, mdl_user_enrolments

Conjunto de tablas II	mdl_feedback, mdl_chat, mdl_assignment, mdl_book, mdl_data, mdl_folder, mdl_glossary, mdl_label, mdl_lesson, mdl_lti, mdl_page, mdl_quest, mdl_resource, mdl_scorm, mdl_survey, mdl_url, mdl_wiki, mdl_workshop.

De igual manera en esta fase una vez identificadas las tablas potenciales para el estudio, se preparó la infraestructura Big Data; se optó por la plataforma de virtualización VirtualBox que ofrece mayores funcionalidades para la gestión de máquinas virtuales. Se procedió a crear un nuevo clúster en el cual se exporta el servicio virtualizado de Apache Hadoop Cloudera, que incluye herramientas como Sqoop, HDFS y Hive. Preparada la infraestructura se transfirió las tablas definidas para el estudio a la infraestructura Big Data como se muestra en la figura 3.

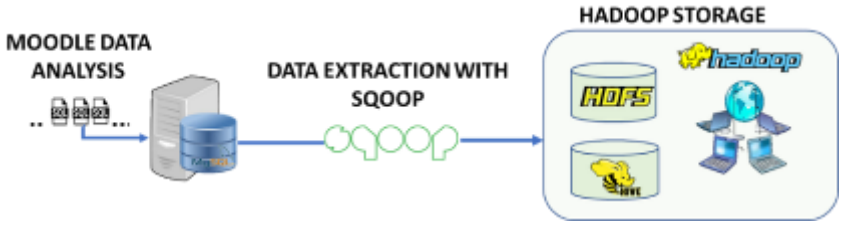

Figura 3. Interaction of various Big Data tools.

En la segunda fase a partir de los datos almacenados en la infraestructura Big data, se realiza la definición de nuevas variables como se muestra en la tabla 3, en este procedimiento es necesario un filtrado de los datos con la finalidad de seleccionar aquellos que son importantes para el proceso de extracción del conocimiento y descartar todo lo innecesario, para ellos se utilizó técnicas como consultas HiveQL y análisis estadístico a los datos almacenados en Hive, cuyos resultados son almacenados en una nueva base de datos.

Tabla 2. Variables predictoras para generación del modelo de predicción.

Variable	Descripción
Curso	Nombre corto del curso
Estudiante	Identificación del estudiante
num_accesos	Número total de vistas de un estudiante en el curso
Num_accesos_unicos	Número de accesos únicos realizados por día.
num_tiempo_curso	Número total de minutos que un estudiante vio el curso.
num_visitas_recurso	Número total de visitas a un módulo recurso (Archivo, carpeta, página, url)

Num_visitas_url	Número de visitas al recurso url.
num_quiz_completados	Número total de cuestionarios completados por un estudiante en un curso
num_post_foro	El número total de foros vistos por un estudiante
num_respuestas_discusiones	Número total de respuestas realizadas por un alumno en el foro de un curso.
num_tareas_entregadas	Número de tareas entregadas
num_vistas_noches	Número total de visitas que un estudiante vio un curso por la noche (7 pm a 5 am)
num_vistas_fin_semana	El número total de vistas que un estudiante vio un curso durante los fines de semana (viernes, sábado y domingo).
num_post_foro	Número de participaciones realizadas en foros.
num_actividades asignadas	Total, de tareas asignadas en el curso (tipo tarea, cuestionarios, foros, tarea externa, etc.)
num_actividades_completadas	Número de actividades completadas en el curso.
num_quiz_completados	Número de cuestionarios realizados
num_tarea_asig_ext	Total, de tareas entregadas con el módulo tarea y herramienta externa Google.
porc_curso_completo	Porcentaje de relación entre tareas entregadas y total de tareas propuestas en el curso.
num_vistas_lunes, num_vistas_martes, num_vistas_miercoles, num_vistas_jueves num_vistas_viernes. num_vistas_sabado num_vistas_domingo	El número total de vistas que un estudiante vio un curso en los diferentes días de la semana.
nota_final	Calificación final de un estudiante en un curso que es A (Aprobado) o P (Reprobado).

Finalmente, en la tercera fase, se aplica la metodología CRISP-DM la cual profundiza en detalle las tareas y actividades a realizar para analizar el comportamiento de los estudiantes y predicción, resultado que sirven de apoyo para la mejora del proceso de enseñanza – aprendizaje virtual.

Una vez computados los datos registrados en el sistema correspondiente a 1731 cursos donde participaron 24556 estudiantes se obtuvieron los registros que corresponden a dos periodos académicos del año 2021 de la modalidad presencial, que debido a la pandemia por el Covid 2019 en dichos periodos las actividades se llevaron a cabo de manera virtual por lo que se tomó en cuenta todas las materias que se imparten por carrera en la Universidad Técnica

de Manabí. Con base a las variables predictoras establecidas se obtuvo un total de 112423 registros para la creación del modelo de 11 facultades e institutos como se observa en la tabla 4.

Tabla 3: Datos computados de cursos por facultades de la UTM del perdido 2021-P1 y 2021-P2 para el modelado.

Nro.	Facultad	Nro. registros	Porcentaje
1	Ciencias administrativas y económicas	19185	17.07
2	Ciencias de la salud	9627	8.56
3	Ciencias humanísticas y sociales	16244	14.45
4	Ciencias informáticas	5264	4.68
5	Ciencias matemáticas físicas y químicas	15729	13.99
6	Ciencias veterinarias	2860	2.54
7	Filosofía letras y ciencias de la educación	14337	12.75
8	Ingeniería agrícola	737	0.66
9	Ingeniería agronómica	665	0.59
10	Instituto de ciencias básicas	13350	11.87
11	Instituto de lenguas	14425	12.83
Total		**112423**	**100**

Una vez obtenidos los 112423 registros para el modelado se realizó una división, destinando el 80% para entrenamiento y el 20% para evaluación como se muestra en la tabla 5.

Tabla 4. Cantidad de datos para el modelado y evaluación.

Destino	%	Cantidad
Entrenamiento	80	89938
Evaluación	20	22485
Total		112423

Como parte de la fase de preparación de los datos, se identificó las variables que presentan valores atípicos, realizando el respectivo tratamiento, de igual manera al analizar la variable rendimiento académico, se evidencio un desbalance en los datos, teniendo un 21.79% de estudiantes reprobados de los 89938 registros para entrenamiento, siendo necesario la aplicación de técnicas de balanceo undersampling y oversampling, que luego de su aplicación se obtuvo tres conjunto de datos para la aplicación de los diversos modelos como se observa en la tabla 6.

Tabla 5. Conjunto de datos para aplicación de algoritmos de clasificación.

Tipo de datos	Cantidad de registros	Descripción
Original	89938	Registros con ambas clases sin balancear
Undersampling	39206	Registros con clase minoritaria balanceada- reprobados
Oversampling	140670	Registros con clase mayoritaria balanceada - aprobados

Resultados y discusión

La ejecución de cada una de las fases de la metodología propuesta, permitió definir la infraestructura Big Data a utilizar para el almacenamiento y procesamiento de grandes volúmenes de datos, así como establecer una vez realizado el análisis de la base de datos de Moodle las principales variables que definen el comportamiento de los estudiantes y además a partir de ellas crear un modelo predictivo del rendimiento académico de los estudiantes con la utilización de algoritmos supervisados. Para la implementación de los modelos propuestos y despliegue se utilizó la herramienta Power Bi, la cual facilita a docentes y demás autoridades la visualización de los resultados que permiten conocer a modo resumen las principales interacciones realizadas en la plataforma que conlleva a la toma de decisiones para la mejora del proceso educativo realizado en entornos virtuales.

Realizada la división, el tratamiento de los datos y seleccionadas las variables de mayor importancia para los modelos, se aplicó los algoritmos de clasificación a los diferentes conjuntos de datos, para identificar con qué conjunto de datos los algoritmos presentan un mejor rendimiento. Se analizaron un total de ocho algoritmos supervisados disponibles en la biblioteca scikit-learn de Python; adicional se aplica el ajuste de los parámetros a los algoritmos Random forest y XGBoost para la obtención de diez modelos, que luego de ser entrenados se evalúan los resultados a través de las métricas de evaluación de modelos de machine learning como son la precisión, sensibilidad, exactitud y puntaje de F1, para cada conjunto de datos como se muestran en la tabla 6.

Tabla 6. Comparación de modelos según parámetros de la matriz de confusión.

Algoritmos	Datos	Precisión	Sensibilidad	Exactitud	Puntaje de F1
XGBoost	Original	0.871	0.970	0.864	0.918
	Undersampling	0.911	0.856	0.821	0.882
	Oversampling	0.910	0.859	0.823	0.884
XGBoost Tunning	Original	0.880	0.971	0.873	0.923
	Undersampling	0.922	0.859	0.832	0.889
	Oversampling	0.919	0.869	0.837	0.893
Catboost	Original	0.893	0.972	0.886	0.931

Algoritmos	Datos	Precisión	Sensibilidad	Exactitud	Puntaje de F1
	Undersampling	0.930	0.859	0.839	0.893
	Oversampling	0.931	0.884	0.858	0.907
Random Forest Classifier	Original	0.893	0.970	0.886	0.930
	Undersampling	0.935	0.855	0.839	0.893
	Oversampling	0.902	0.957	0.885	0.929
Random Forest Classifier Tunnig	Original	0.847	0.982	0.846	0.909
	Undersampling	0.903	0.863	0.819	0.882
	Oversampling	0.898	0.876	0.825	0.887
Decision Tree Classifier	Original	0.884	0.883	0.826	0.889
	Undersampling	0.915	0.742	0.743	0.819
	Oversampling	0.889	0.889	0.826	0.889
SVC - Linear Kernel	Original	0.820	0.989	0.822	0.897
	Undersampling	0.861	0.845	0.772	0.853
	Oversampling	0.868	0.852	0.782	0.860
k-vecino	Original	0.834	0.948	0.811	0.887
	Undersampling	0.856	0.678	0.658	0.757
	Oversampling	0.849	0.690	0.661	0.762
Naive Bayes	Original	0.858	0.864	0.781	0.861
	Undersampling	0.878	0.727	0.706	0.795
	Oversampling	0.875	0.737	0.711	0.800
Logic Regression	Original	0.819	0.987	0.818	0.895
	Undersampling	0.859	0.821	0.754	0.839
	Oversampling	0.863	0.803	0.746	0.832

En los resultados obtenidos para cada algoritmo y conjunto de datos como se observa en la tabla 6, considerando la métrica “precisión” los modelos presentan mejor rendimiento con los datos balanceados al aplicar la técnica undersampling y con la métrica “sensibilidad” con el conjunto de datos no balanceados presentan mejor rendimiento; al evaluar los algoritmos con la métrica “puntaje de f1” que combina el rendimiento de “precisión” y “sensibilidad” se obtiene mejor resultado con los datos originales.

Considerando el análisis realizado a los resultados y teniendo en cuenta el objetivo de la investigación; para la selección del algoritmo aplicado al proyecto con el cual se obtiene una mejor calidad en los pronósticos se selecciona la métrica "precisión". De acuerdo a los resultados presentados en la tabla 6, el algoritmo que presenta mejor rendimiento es el "Random forest" con 94% de precisión, 86% de sensibilidad, seguido del "CatBoost" con un 93% de precisión y el XGBoost con 92%.

Seleccionado el modelo "Random forest", se analizó la matriz de confusión que muestra el desempeño del algoritmo de clasificación donde se obtiene como verdaderos negativos 3812 registros y 15050 verdaderos positivos, como se observa en la figura 4.

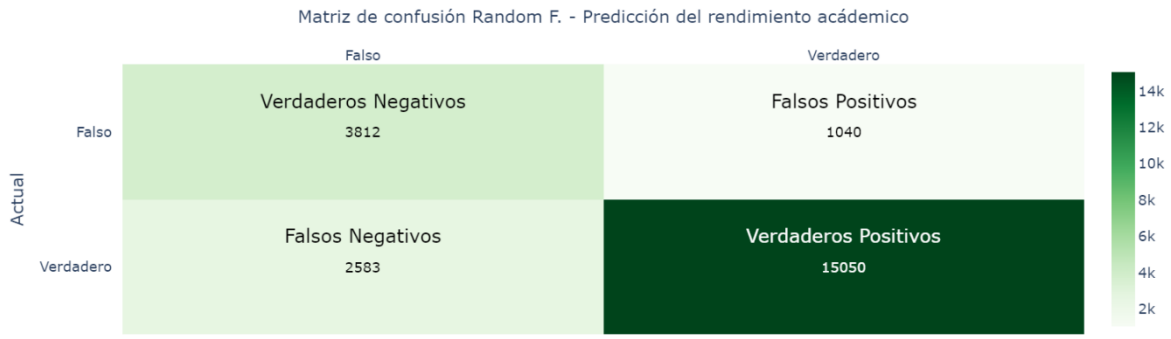

Figura 4. Matriz de confusión del Random forest.

Con base a la matriz de confusión se determina la precisión de 0.935 para el algoritmo Random forest, de acuerdo a la siguiente formula:

$$Precisión = \frac{VP}{VP + FP}$$

$$Precisión = \frac{15050}{15050 + 1040} = 0.935$$

Otra forma de evaluar los algoritmos es compararlos gráficamente trazando los de mayor rendimiento, para ello se utilizó la curva ROC en la que muestra la tasa de falsos positivos y falsos negativos, revelando nuevamente que el Random forest obtiene un AUC de 0.820 mayor al CatBoost y XGBoost, y de igual forma en la curva Precisión-Recall el Random forest cuenta con mayor rendimiento como se muestra en la figura 5.

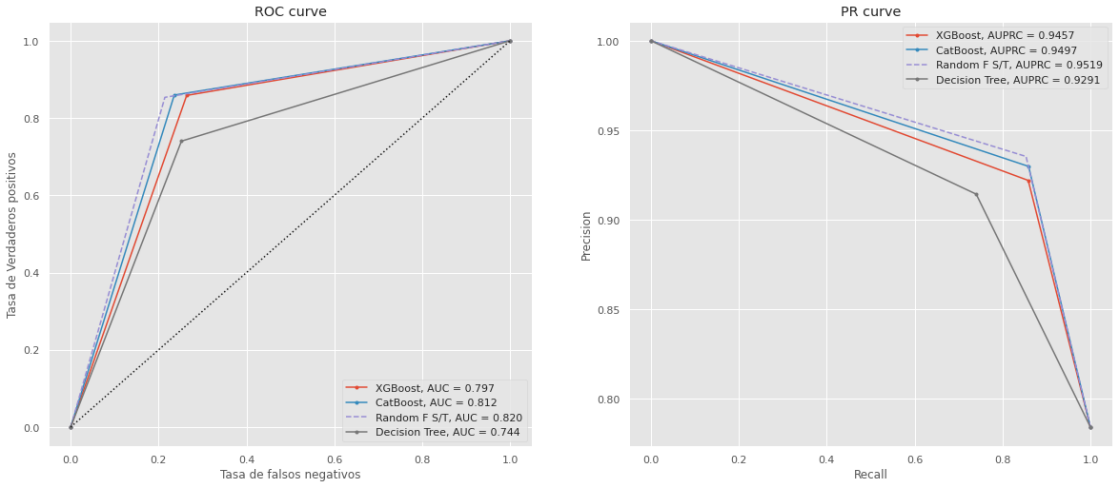

Figura 5. Curva ROC y PR.

Para el de despliegue y la visualización de los resultados se utiliza la herramienta Power Bi, la cual permite tratar con un gran volumen de datos, adicional en ella se realizado la ejecución del script con el modelo Random Forest y se realiza la predicción a partir de un nuevo dataset proporcionado acorde a la estructura establecida de los datos, obteniendo informes de predicción del rendimiento académico de los estudiantes, como se observa en las figuras 6 y 7.

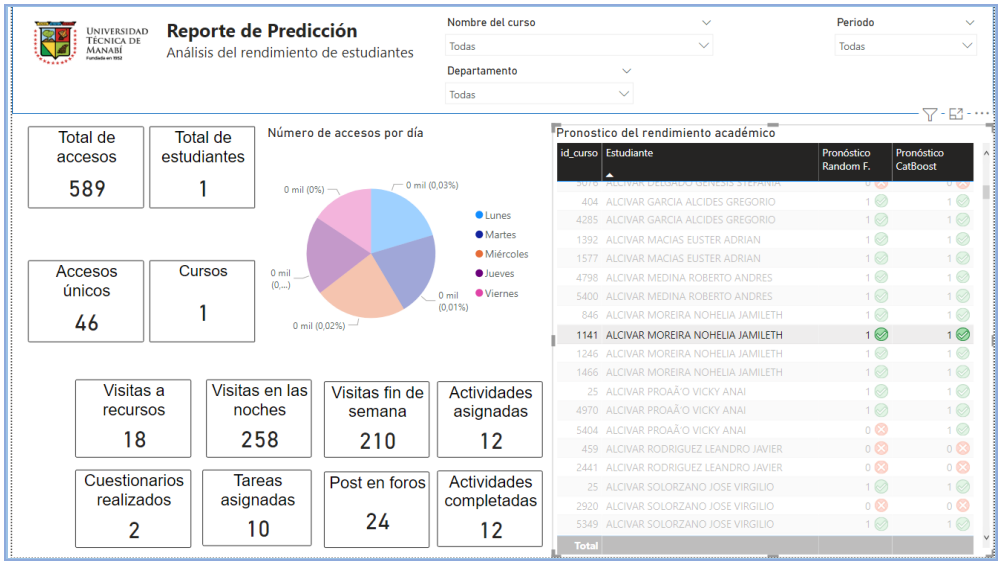

Figura 6. Predicción estudiante que "Aprueba un curso".

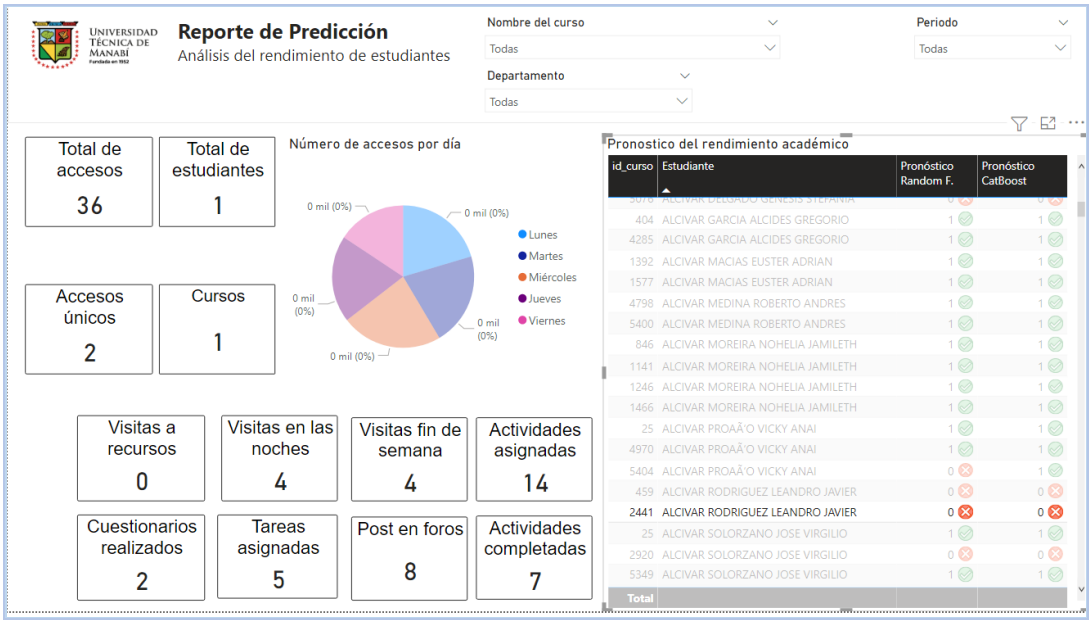

Figura 7. Predicción estudiante "Reprueba un curso".

Este estudio reveló que el uso de Big Data y tecnologías como Hadoop, machine learning y Power BI es crucial para el almacenamiento, procesamiento y análisis de grandes volúmenes de datos académicos. Estos resultados coinciden con investigaciones anteriores que destacan los beneficios del uso de técnicas y herramientas de Big Data en entornos universitarios (Amaya-Amaya et al., 2020; Baig & Jabeen, 2016). Las variables predictoras son clave para predecir el rendimiento académico de los estudiantes, como el número de actividades completadas, accesos a la plataforma, participación en foros y calificaciones. Estos hallazgos se respaldan con estudios similares que utilizan variables similares y demuestran la relevancia de la forma en que los estudiantes interactúan con la plataforma Moodle (Akçapınar, 2016; Bogarín Vega et al., 2015; J. López et al., 2020; Menacho, 2020; Roldan & Castro, 2016). El desarrollo de modelos predictivos utilizando algoritmos supervisados, y en caso de los tres mejores modelos Random Forest, Catboost y XGboost lograron un rendimiento superior al 90% en la predicción del rendimiento académico de los estudiantes superando en algunos casos los resultados de otros estudios (Bognar & Fauszt, 2020; Felix et al., 2016; Liang et al., 2016; Rico Páez et al., 2019; Roldan & Castro, 2016; Saiz Manzanares et al., 2018). Esto demuestra la capacidad de obtener predicciones confiables a partir de grandes conjuntos de datos y proporciona evidencia de un rendimiento similar a investigaciones previas en el campo.

Conclusiones

El análisis de la información generada por la interacción de los estudiantes de pregrado en la plataforma Moodle, mediante mecanismo de Big Data, permite conocer el comportamiento de los usuarios en cursos virtuales, y como sus acciones inciden en su aprendizaje; así como a partir del entrenamiento y evaluación de los diversos algoritmos de machine learnig, obtener un modelo de predicción del rendimiento académico. De acuerdo al estudio realizado y teniendo en cuenta el gran volumen de datos generado en la plataforma Moodle, se establece que una de las herramienta a emplear para el almacenamiento, procesamiento y análisis de los datos académicos, es la distribución de Hadoop Cloudera, la cual integra tecnologías como HDFS, Sqoop y Hive. La aplicación de la técnica de clasificación de machine learning, para la obtención del modelo de predicción del rendimiento académico de los estudiantes y la herramienta Power Bi para la presentación de los resultados. Se definieron que las variables que permiten predecir el rendimiento académico son el número de actividades completadas, números de accesos totales y únicos, número de quiz completados, número de visitas en fines de semana, visitas en las noches, número de tareas asignadas, número de participaciones en foros, visitas a recursos, números de actividades asignadas, porcentaje del curso completo y calificación.

A partir del entrenamiento y modelado de ocho algoritmos supervisados se obtuvo el modelo Random Forest aplicado a los datos balanceados con la técnica undersampling, el cual presenta mejores. Se validó los modelos obtenidos con la métrica “precisión” para predecir el rendimiento académico de los estudiantes y clasificarlos como “Aprobado” o “Reprobado” teniendo que los mejores resultados de la predicción se obtuvieron con los algoritmos Random Forest con un 94%, seguido el CatBoost con 93% y el XGBoost con el 92% de precisión. El algoritmo Random Forest logra un área bajo la ROC y PR, en el conjunto de datos de validación de 82% y 95% respectivamente, lo cual indica un buen desempeño en comparación a los demás algoritmos, aunque se selecciona este algoritmo en este estudio, se determina que los algoritmos CatBoost y XGBoost presentan un rendimiento similar, pudiendo utilizar cualquiera de ellos al no presentar diferencias significativas.

Conflictos de intereses

Los autores no poseen conflictos de intereses.

Contribución de los autores

1. Conceptualización: Yandry José Olarte Sancán.
2. Curación de datos: Marely del Rosario Cruz Felipe

3. Análisis formal: Juan Andrés Mejía Almenaba
4. Investigación: Yandry José Olarte Sancán, Jenmer Maricela Pinargote Ortega
5. Metodología: Yandry José Olarte Sancán. Marely del Rosario Cruz Felipe
6. Administración del proyecto: Yandry José Olarte Sancán,
7. Software: Yandry José Olarte Sancán.
8. Supervisión: Marely del Rosario Cruz Felipe
9. Validación: Juan Andrés Mejía Almenaba, Jenmer Maricela Pinargote Ortega
10. Visualización: Juan Andrés Mejía Almenaba
11. Redacción – Yandry José Olarte Sancán.
12. Redacción – revisión y edición: Yandry José Olarte Sancán., Marely del Rosario Cruz Felipe

Financiamiento

La investigación no requirió fuente de financiamiento externo.

Referencias

Akçapınar, G. (2016). *Predicting students' approaches to learning based on moodle logs*. *1*(July), 2347–2352. https://doi.org/10.21125/edulearn.2016.1473

Amaya-Amaya, A., Huerta-Castro, F., & Flores-Rodríguez, C. O. (2020). Big Data, una estratégica para evitar la deserción escolar en las IES. *Revista Iberoamericana de Educación Superior*, *11*(31), 166–178. https://doi.org/10.22201/iisue.20072872e.2020.31.712

Baig, A. R., & Jabeen, H. (2016). Big Data Analytics for Behavior Monitoring of Students. *Procedia Computer Science*, *82*(March), 43–48. https://doi.org/10.1016/j.procs.2016.04.007

Belmonte, J., Sánchez, S., & Guerrero, A. J. (2019). Consideraciones sobre el B-learning en el proceso de enseñanza y aprendizaje. *Universidad&Ciencia*, *8*(2), 24–39. http://revistas.unica.cu/index.php/uciencia/article/view/1239

Bogarín Vega, A., Romero Morales, C., & Cerezo Menéndez, R. (2015). Aplicando minería de datos para descubrir rutas de aprendizaje frecuentes en Moodle. *EDMETIC*, *5*(1), 73. https://doi.org/10.21071/edmetic.v5i1.4017

Bognar, L., & Fauszt, T. (2020). Different learning predictors and their effects for moodle machine learning models. *11th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2020 - Proceedings, September 2020*, 405–410. https://doi.org/10.1109/CogInfoCom50765.2020.9237894

Felix, I., Ambrosio, A. P., Duilio, J., & Simões, E. (2016). *Predicting Student Outcome in Moodle*.

Gámez, F. I. L., Rodríguez, M. R., & Torres, L. E. S. (2018). Uso y aplicación de las TIC en el proceso de enseñanza-aprendizaje. *Revista Científica de FAREM-Estelí*, *25*, 16–30. https://doi.org/10.5377/FAREM.V0I25.5667

Hidalgo Cajo, B. G. (2018). data mining in Learning Management Systems in University Education. In *Campus Virtuales* (Vol. 7, Issue 2). Campus Virtuales. www.revistacampusvirtuales.es

Liang, J., Yang, J., Wu, Y., Li, C., & Zheng, L. (2016). Big data application in education: Dropout prediction in edx MOOCs. *Proceedings - 2016 IEEE 2nd International Conference on Multimedia Big Data, BigMM 2016*, 440–443. https://doi.org/10.1109/BigMM.2016.70

López, E., Cobos, D., Martín, A., Molina, L., & Jaén, A. (2018). Uso de Moodle por alumnos y rendimiento académico. *Experiencias Pedagógicas e Innovación Educativa. Aportaciones Desde La Praxis Docente e Investigadora*, 1634–1642. https://rio.upo.es/xmlui/handle/10433/6411

López, J., Lara, J. A., & Romero, C. (2020). Towards portability of models for predicting students' final performance in university courses starting from moodle logs. *Applied Sciences (Switzerland)*, *10*(1). https://doi.org/10.3390/app10010354

Menacho, C. (2020). Técnicas de minería de datos aplicadas a la plataforma educativa Moodle. *Revista Tierra Nuestra*, *14*(1), 137–146. https://190.119.243.75/index.php/tnu/article/download/1509/1911

Menacho Chiok, C. H. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Anales Científicos*, *78*(1), 26. https://doi.org/10.21704/AC.V78I1.811

Miranda, M. A., & Guzmán, J. (2017). Análisis de la deserción de estudiantes universitarios usando técnicas de minería de datos. *Formacion Universitaria*, *10*(3), 61–68. https://doi.org/10.4067/S0718-50062017000300007

Muñoz-Gea, J. P., Pérez de la Cruz, F. J., Busquier Sáez, S., Silva Pérez, M. M., & Angosto Hernández, C. (2016). Interacción de los estudiantes con las actividades de Moodle: un estudio basado en Web Mining / Student Interaction with Moodle Activities: a Study Based on Web Mining. *Revista Internacional de Tecnología, Ciencia y Sociedad*, *5*(1), 19–28. https://doi.org/10.37467/gka-revtechno.v5.453

Rico Páez, A., Gaytán Ramírez, N. D., & Sánchez Guzmán, D. (2019). Construcción e implementación de un modelo para predecir el rendimiento académico de estudiantes universitarios mediante el algoritmo Naïve Bayes. *Diálogos Sobre Educación*, *10*(19). https://doi.org/10.32870/dse.v0i19.509

Roldan, A., & Castro, J. (2016). *Mecanismos de análisis BigData para la caracterización de la actividad docente en un Campus Virtual Moodle*. http://uvadoc.uva.es/bitstream/handle/10324/17475/TFM-G547.pdf?sequence=1&isAllowed=y

Saiz Manzanares, M. C., Marticorena Sánchez, R., Arnaiz González, Á., Escolar Llamazares, M. del C., & Queiruga

Dios, M. Á. (2018). Detección del alumno en riesgo en titulaciones de Ciencias de la Salud: aplicación de técnicas de Learning Analytics. *European Journal of Investigation in Health, Psychology and Education*, *8*(3), 129. https://doi.org/10.30552/ejihpe.v8i3.273

SGA, & UTM. (2021). *Total de estudiantes inscritos/reprobados agrupados por asignatura departamental*. https://app.utm.edu.ec/sga/

Yang, Y., Hooshyar, D., Pedaste, M., Wang, M., Huang, Y. M., & Lim, H. (2020). Predicting course achievement of university students based on their procrastination behaviour on Moodle. *Soft Computing*, *24*(24), 18777–18793. https://doi.org/10.1007/s00500-020-05110-4