

Artículo

Hacia la Predicción de la Abandono Estudiantil en Carreras Universitarias Uso de diferentes técnicas de aprendizaje automático

Janka Kabathova y Martin Drlik *

Departamento de Informática, Facultad de Ciencias Naturales, Universidad Constantino el Filósofo en Nitra, Tr. A.
Hlinku 1, 949 01 Nitra, Eslovaquia; janka.kabathova@ukf.sk * Correspondencia: mdrlik@ukf.sk

Aplicación destacada: El modelo encontrado con los mejores valores de las métricas de rendimiento, encontrado como resultado de comparar varios clasificadores de aprendizaje automático, puede identificar a los estudiantes en riesgo a pesar de que el conjunto de datos educativos obtenidos durante la ejecución del curso es escaso. Como resultado, se puede aplicar a tiempo una forma adecuada de intervención a nivel de curso de e-learning individual.

Resumen: La predicción temprana y precisa de la deserción de los estudiantes en función de los datos educativos disponibles pertenece al campo de investigación generalizado del campo de investigación de las analíticas de aprendizaje. A pesar de la cantidad de investigaciones ya realizadas, el progreso no es significativo y persiste en todos los niveles de datos educativos. Aunque ya se han investigado varias características, todavía queda una pregunta abierta, qué características pueden considerarse apropiadas para diferentes clasificadores de aprendizaje automático aplicados al conjunto escaso típico de datos educativos a nivel de curso de aprendizaje electrónico. Por lo tanto, el objetivo principal de la investigación es enfatizar la importancia de la comprensión de datos, la fase de recopilación de datos, enfatizar las limitaciones de los conjuntos de datos educativos disponibles, comparar el rendimiento de varios clasificadores de aprendizaje automático y mostrar que también un conjunto limitado de datos. Las características, que están disponibles para los profesores en el curso de aprendizaje electrónico, pueden predecir la deserción de los estudiantes con suficiente precisión si se consideran minuciosamente las métricas de rendimiento. Se analizaron los datos recogidos de cuatro años académicos. Las características seleccionadas en este estudio demostraron ser aplicables en la predicción de los que completaron y no completaron el curso. La precisión de la predicción varió entre el 77 y el 93 % en datos no vistos del próximo año académico. Además de las métricas de rendimiento de uso frecuente, se analizó la comparación de la homogeneidad de los clasificadores de aprendizaje automático para superar el impacto del tamaño limitado del conjunto de datos en los altos valores obtenidos de las métricas de rendimiento. Los resultados mostraron que varios algoritmos de aprendizaje automático podrían aplicarse con éxito a. Simultáneamente, las métricas de desempeño de la clasificación se deben considerar detenidamente antes de decidir implementar el mejor modelo de clasificación de desempeño para predecir posibles casos de deserción y diseñar mecanismos de intervención beneficiosos.

Palabras clave: analítica de aprendizaje; minería de datos educativos; aprendizaje automático; predicción de abandono



Cita: Kabathova, J.; Drlik, M.

Hacia la predicción de la deserción
de los estudiantes en cursos universitarios
utilizando diferentes técnicas de aprendizaje
automático. *aplicación ciencia* **2021**, *11*, 3130.
<https://doi.org/10.3390/app11073130>

Editor Académico: Juan
A. Gómez-Pulido

Recibido: 10 febrero 2021

Aceptado: 30 de marzo de 2021

Publicado: 1 abril 2021

Nota del editor: MDPI se mantiene neutral
con respecto a reclamos jurisdiccionales en
mapas publicados y afiliación institucional
aciones.



Copyright: © 2021 por los autores.

Licenciario MDPI, Basilea, Suiza.

Este artículo es un artículo de acceso abierto
distribuido bajo los términos y

condiciones de Creative Commons

Licencia de atribución (CC BY) ([https://creativecommons.org/licenses/by/](https://creativecommons.org/licenses/by/4.0/)

4.0/).

1. Introducción

Hoy en día, los entornos virtuales de aprendizaje (VLE, por sus siglas en inglés) representan plataformas de educación en línea maduras que permiten a los docentes brindar cursos que contienen recursos administrados minuciosamente y una miríada de diferentes actividades atractivas. Los VLE permiten crear varios juegos educativos, cuestionarios, pruebas de aprendizaje electrónico o preparar materiales multimedia para atraer a los estudiantes.

Teniendo en cuenta el progreso tecnológico y la creciente demanda de educación en línea, los VLE han percibido un rápido desarrollo en los últimos años [1]. Los desarrolladores y proveedores de VLE han aprovechado al máximo Internet y las tecnologías web actuales y operan sistemas educativos modernos para mejorar el conocimiento y las habilidades de los estudiantes. Los enfoques y metodologías educativas adaptadas a las funcionalidades de las plataformas virtuales eliminan las limitaciones de los cursos tradicionales tomados en las aulas y brindan una mayor flexibilidad en términos de dónde y cuándo tomar los cursos en línea.

Las herramientas computacionales modernas para el análisis de datos, la estandarización de formatos para el almacenamiento de datos y el aumento de la potencia informática permiten que las preguntas de investigación se exploren ampliamente utilizando diversos enfoques basados en datos. Dado que los entornos virtuales de aprendizaje están diseñados para capturar y almacenar la mayor cantidad posible de datos sobre la interacción del usuario con el sistema, no sorprende que estos datos se analicen utilizando diferentes técnicas de descubrimiento de conocimientos para identificar los problemas relacionados con el aprendizaje de manera temprana y sugerir la solución más adecuada. Como resultado, diferentes técnicas de análisis de aprendizaje y minería de datos educativos ya han ayudado a analizar y comprender parcialmente el comportamiento de los estudiantes, detectar errores de aprendizaje frecuentes, recomendar rutas de aprendizaje, mejorar las pruebas y medir el desempeño de los estudiantes.

A pesar del rápido desarrollo y los éxitos de la investigación basada en datos, las instituciones educativas aún enfrentan muchas dificultades principales, como la interacción efectiva de las partes interesadas de EVA y la retroalimentación inmediata, que no son adecuadas, son insuficientes o se retrasan. Este hecho influye posteriormente en el rendimiento educativo de los estudiantes y en su resultado final [2].

La deserción de los estudiantes plantea uno de esos problemas significativos dentro de los enfoques contemporáneos de aprendizaje electrónico [3]. La deserción tiene un impacto negativo en las organizaciones educativas, así como en las partes interesadas. Aunque ha sido el foco de muchas investigaciones, las tasas de abandono del e-learning generalmente tienden a ser más altas que las de la educación presencial [4].

Por lo tanto, los investigadores aplican minuciosamente diferentes métodos de descubrimiento de conocimiento y examinan el comportamiento del estudiante, sus rasgos característicos, así como el impacto del entorno externo en el proceso de aprendizaje y los resultados para minimizar la deserción, crear una estrategia de intervención efectiva y proporcionar retroalimentación inmediata [5].

Los datos educativos utilizados en los modelos de predicción de la deserción provienen de diferentes fuentes, tienen diferente tamaño y calidad. Los datos educativos a menudo se almacenan en sistemas de bases de datos. Puede consistir en información sobre las actividades de los estudiantes y proporcionar a los usuarios finales estadísticas de todo el curso, brindando una descripción general agregada de la actividad de todas las partes interesadas. Además, también se dispone información resumida sobre las notas parciales y finales sobre el desempeño de cada estudiante. Finalmente, a menudo se utilizan registros de eventos, que rastrean indirectamente la interacción de cada parte interesada con la plataforma [6].

La disponibilidad de los recursos de datos educativos para la investigación basada en datos difiere según los privilegios del investigador y el alcance del problema de investigación. Mientras que el investigador puede tener una cantidad considerable de datos disponibles en forma de registros, la cantidad de datos que describen los logros de los estudiantes en el nivel del curso es limitada.

La combinación de estos recursos no se correlaciona automáticamente con la mayor precisión de las predicciones finales de la deserción de los estudiantes porque los datos deben agregarse con frecuencia.

Por lo tanto, el objetivo principal del artículo es enfatizar la importancia de la comprensión de datos, la fase de recopilación de datos, enfatizar las limitaciones de los conjuntos de datos educativos disponibles, comparar el rendimiento de varios clasificadores de aprendizaje automático y mostrar que también un conjunto limitado de Las características, que están disponibles para los profesores en el curso de aprendizaje electrónico, pueden predecir la deserción de los estudiantes con suficiente precisión si se consideran minuciosamente las métricas de rendimiento.

Aunque el tema principal del documento es bastante frecuente, el documento describe un estudio de caso, que enfatiza la necesidad de seguir una metodología de descubrimiento de conocimiento bien conocida incluso en el caso de clasificadores de uso común y considerar no solo un conjunto limitado de métricas de rendimiento. El enfoque aplicado tiene el potencial de predecir la deserción del estudiante en el nivel del curso si la estructura del curso no ha cambiado significativamente y el curso se ha reabierto durante varios años.

El trabajo de investigación está organizado de la siguiente manera. En primer lugar, se presenta el resumen de trabajos relacionados enfocados en predicciones efectivas del éxito de los estudiantes utilizando técnicas de aprendizaje automático en el dominio de analítica de aprendizaje. Las fases individuales de la metodología CRISP-DM, que gestiona el proceso de descubrimiento de conocimiento, se introducen gradualmente, a partir de la Sección 3. Esta sección proporciona información sobre las principales características del conjunto de datos investigado. Posteriormente, se describen en detalle las fases de comprensión y preprocesamiento de los datos. La sección 4 trata sobre los modelos de predicción de la deserción del estudiante, en los que se han aplicado diferentes clasificadores de aprendizaje automático. Sección 5

se refiere a los resultados obtenidos y la evaluación de los modelos basados en diferentes medidas de rendimiento. Además, el rendimiento de los clasificadores de aprendizaje automático se compara mediante la prueba de McNemar. Finalmente, la Sección 6 resume los resultados de la investigación, discute las limitaciones de la investigación y propone su posible dirección futura.

2. Trabajos relacionados

En estudios previos se establecieron varias definiciones de abandono de los estudiantes. La definición más frecuente se centra en si los estudiantes seguirán activos hasta la última semana o si la semana actual es la última semana en la que los estudiantes estarán activos. La identificación temprana de los estudiantes en riesgo de abandono juega un papel importante para reducir el problema y permitir que se aborden las condiciones requeridas.

Las consideraciones de tiempo son, por lo tanto, significativas para el problema de la deserción. Como resultado, algunos análisis han demostrado que el 75% de los abandonos se producen en las primeras semanas [7]. La predicción de la deserción a menudo se considera un problema de predicción de series temporales o un problema de etiquetado de secuencias [8]. Estos pueden corresponder con el estado final de los estudiantes [9]. Por otro lado, la dimensión temporal se puede incluir en la predicción de abandono indirectamente utilizando las características de entrada disponibles en una ventana de tiempo particular, que permiten seleccionar una forma adecuada de intervención [10].

La predicción del éxito o fracaso del aprendizaje del estudiante pertenece a uno de los temas más investigados de las disciplinas de minería de datos educativos (EDM) y análisis de aprendizaje (LA) [3]. EDM se ocupa del análisis de datos relacionados con el estudio para comprender el comportamiento de los estudiantes. Estas técnicas generalmente se aplican para proporcionar entornos de aprendizaje más efectivos al revelar información útil para modificar la estructura de los cursos o respaldar la predicción del rendimiento y el comportamiento de los estudiantes [11]. Por otro lado, el análisis del aprendizaje se ocupa de la medición, recopilación, análisis y presentación de informes de los datos de los estudiantes y sus antecedentes para comprender y mejorar el aprendizaje y los entornos en los que tiene lugar [12]. El núcleo de los enfoques de predicción actuales generalmente se basa en los métodos EDM y LA. Uno de los temas de investigación más candentes en el análisis del aprendizaje ha sido la predicción de la probabilidad de que los estudiantes completen o reprobren un curso, especialmente desde las primeras semanas [13], así como EDM [14]. Una vez que se dispone de una predicción fiable del rendimiento, se puede utilizar en muchos contextos para identificar a los estudiantes débiles y proporcionar retroalimentación a los estudiantes o predecir el fracaso de los estudiantes [15].

Una predicción exitosa depende del uso de datos confiables, precisos y significativos [9]. El tamaño del conjunto de datos es un escollo típico de muchos estudios de investigación publicados en tecnología educativa, análisis de aprendizaje e investigación de minería de datos educativos [16]. Es un desafío alcanzar el volumen de datos educativos que satisfaría completamente los requisitos de aprendizaje automático además de usar registros [8]. Por otro lado, la alta granularidad y la naturaleza técnica de los registros VLE a menudo pierden la semántica necesaria, lo que puede causar problemas sobre cómo deben interpretarse los hallazgos desde el punto de vista educativo [17].

Hämäläinen y Vinni indicaron en su revisión de clasificadores para EDM que los conjuntos de datos que se utilizan a menudo en la minería de datos educativos son relativamente pequeños. Suelen constar de 200 registros en promedio en el nivel del curso [18]. Muchos experimentos analizados confirmaron esta afirmación. Se utilizó un volumen de datos similar para evaluar las técnicas de aprendizaje automático en el nivel de curso VLE individual [18].

Como afirmaron Ifenthaler y Yau [16], más datos educativos no siempre han llevado a mejores decisiones educativas. Los mismos autores concluyeron que la confiabilidad y validez de los datos y su análisis preciso y sin sesgos son fundamentales para generar conocimientos sumativos, en tiempo real o formativos, y predictivos o prescriptivos útiles para el aprendizaje y la enseñanza.

Lang et al. [19] afirmaron que los métodos de clasificación y regresión, la red neuronal, las redes bayesianas, las máquinas de vectores de soporte, la regresión logística y la regresión lineal podrían resolver el problema de predicción del rendimiento de los estudiantes. Estos modelos a menudo se denominan modelos de caja negra que son difíciles de entender e interpretar. Todos ellos también dependen en gran medida de la extracción de características. La extracción de características se refiere a la compilación de un subconjunto de características predictivas únicas para el problema predictivo en el modelado, también conocido como atributo.

selección. El proceso ayuda a identificar atributos relevantes en el conjunto de datos que contribuyen a la precisión del modelo de predicción, por ejemplo, por la última actividad en el curso correspondiente del estudiante para predecir un abandono [4]. Las estrategias efectivas para un tipo de conjuntos de datos pueden no ser suficientes para otro. En este caso, a menudo es necesario desarrollar manualmente nuevas estrategias para extraer características [6].

Por otro lado, los árboles de decisión, clasificador basado en reglas, pertenecen a los modelos de caja blanca que son más comprensibles y fáciles de interpretar porque exponen el proceso de razonamiento que subyace a las predicciones. Otra posibilidad es utilizar algoritmos de agrupamiento o minería de reglas de asociación. El análisis de correlación de las calificaciones obtenidas de los cursos y los atributos que definen los créditos obtenidos por los estudiantes y sus calificaciones promedio también puede ser valioso [19].

Ifenthaler et al. [16] realizó una revisión de la dirección de investigación contemporánea de LA. Distinguen diferentes conjuntos de predictores de la deserción de los estudiantes. El primer conjunto de predictores se basa en datos de comportamiento en línea, principalmente archivos de registro y datos de seguimiento. En este contexto, uno de los desafíos difíciles es determinar cómo involucrar a los estudiantes para que participen en el aprendizaje en línea [20]. El segundo conjunto de predictores se basa en la información de antecedentes de los estudiantes, como la demografía, el estado socioeconómico, la experiencia académica previa y el rendimiento [7]. Muchos estudios también suelen examinar los registros relacionados con los estudiantes, por ejemplo, la edad, el sexo, el estado, porque están fácilmente disponibles en los sistemas de información de cada universidad. Está comprobado que atributos como la calificación promedio de los estudiantes o la calificación promedio obtenida en un curso afectan positivamente la predicción [21].

El análisis de los registros de los estudiantes junto con algunos otros predictores, como la cantidad de publicaciones de discusión, la cantidad de vistas del foro, la cantidad de vistas de la prueba, la cantidad de vistas del módulo, la cantidad de días activos y las actividades de las redes sociales, tenían como objetivo predecir la deserción mediante la implementación de un árbol de decisión [22].

De manera similar a la investigación presentada en este artículo, Kennedy et al. [23] consideró los tipos populares de variables, como envíos de tareas, cambios de tareas (es decir, la cantidad de veces que un alumno pasa de una tarea a otra), días activos y puntos totales. Llegaron a la conclusión de que estas características podrían predecir el éxito de los estudiantes. Sin embargo, no discutieron en detalle el papel del conjunto de datos más pequeño y las métricas de rendimiento.

Skalka et al. señaló que otras actividades especiales, como tareas de programación, también se pueden considerar para predecir el puntaje final en el curso de manera efectiva [12]. Las asignaciones también han sido analizadas en [24,25]. Ambos estudios confirmaron que las tareas podrían usarse con éxito para predecir los resultados de aprendizaje de los estudiantes.

Considerar el papel de las publicaciones en los foros como predictores también es un tema importante de investigación. Pueden ser una fuente adecuada de elementos predictivos (p. ej., el número de publicaciones, el número de comentarios, el número de votos positivos, etc.) que normalmente no están disponibles en los cursos presenciales [1].

Herodotou et al. [26] demostraron que los cursos en los que los profesores utilizaron métodos predictivos produjeron al menos un 15 % más de rendimiento que las clases sin su uso [1]. Sin embargo, es un reto llegar a este estado porque cuantos menos datos hay disponibles, menos precisa es la predicción que se puede obtener.

Además, los investigadores están tratando de encontrar características adicionales que puedan contribuir a una comprensión más profunda de los estudiantes. En general, todos estos enfoques produjeron resultados aceptables, lo que permitió la predicción del desempeño de los estudiantes para lograr una precisión general que varió del 65% al 90% [4]. Optimizar los resultados obtenidos por las predicciones es un desafío considerable porque los diferentes algoritmos suelen presentar una gran variación en las tasas de rendimiento dependiendo de la combinación de varias características (por ejemplo, equilibrio entre clases, cantidad de datos, variables de entrada y otras) [1].

3. Preparación de datos

El propósito del proceso de descubrimiento de conocimiento es extraer y transformar la información en una estructura comprensible para su uso posterior. La minería de datos es parte de este proceso, que implica la integración de técnicas como estadísticas, aprendizaje automático,

reconocimiento de patrones, visualización de datos o análisis de datos [27]. El proceso suele ser iterativo y se centra en descubrir relaciones e información valiosas mediante el análisis de muchas variables.

El proceso de descubrimiento de conocimiento se ha mejorado continuamente y se ha transformado en varias variaciones prácticas. Uno de los más utilizados es el Proceso Estándar Intersectorial para la Minería de Datos, conocido como CRISP-DM. Dado que la creación de modelos predictivos en este documento es una tarea típica de un problema de descubrimiento de conocimiento, también fue impulsada por la metodología CRISP-DM [17].

3.1. Comprensión empresarial

Esta fase inicial de CRISP-DM se centra en comprender el problema y los requisitos desde una perspectiva más amplia. Define los objetivos de análisis y los convierte en una definición de problema de minería de datos y un plan para alcanzar los objetivos.

Los datos utilizados en esta investigación provienen del curso de e-learning de LMS Moodle, el cual se ha impartido durante cuatro años académicos. El curso contenía muchos recursos y actividades interactivas, que involucran a los estudiantes para que sean más activos y reciban conocimientos y habilidades, principalmente basados en el aprendizaje.

La pregunta natural es cómo los puntos obtenidos de estas actividades y el comportamiento de los estudiantes durante el trimestre pueden contribuir a la predicción efectiva del éxito o fracaso de los estudiantes.

Hay muchos enfoques más o menos basados en datos, que pueden proporcionar respuestas parciales. La investigación pedagógica más convencional supone la implementación de fases definidas con precisión, que incluyen la prueba previa, la prueba posterior y la evaluación de hipótesis para identificar predictores adecuados. El tamaño del conjunto de datos se considera restrictivo en este caso, pero se compensa con las pruebas estadísticas elegidas apropiadamente. Las opciones limitadas para la automatización de los pasos individuales de este proceso dificultan su uso más amplio por parte de las partes interesadas de LMS.

Por lo tanto, la investigación educativa contemporánea pone esperanza en los enfoques basados en datos, que prometen una automatización parcial del proceso predictivo, una interpretación potencialmente más sencilla, así como una selección efectiva del tipo de intervención adecuado para un grupo más amplio de partes interesadas.

Aunque el LMS Moodle ya ha implementado su propio modelo central de análisis de aprendizaje basado en el comportamiento de los estudiantes en diferentes actividades del curso, no considera los resultados parciales de los estudiantes y las partes interesadas no lo usan con frecuencia.

Por lo tanto, se aplicaron diferentes algoritmos de aprendizaje automático en este estudio de caso para obtener una predicción precisa, ya que todavía no existe una forma específica de presentar la mejor solución para predecir el rendimiento de los estudiantes en la educación en línea o identificar a los estudiantes que tienen más probabilidades de reprobado el curso.

3.2. Comprensión de datos

La segunda etapa consiste en adquirir la información o acceder a los datos necesarios para resolver el problema. Esta recopilación inicial incluye procedimientos para la recuperación de datos, descripción de datos y verificación de la calidad y precisión de los datos [27].

Los datos educativos extraídos de las bases de datos de VLE se pueden usar para encontrar atributos clave o dependencias entre cursos. Cada VLE ofrece datos adecuados para las tareas de descubrimiento de conocimiento. El conocimiento obtenido de los registros y los resultados del estudio puede mejorar los perfiles de los estudiantes [28]. Cuanto más se sepa sobre el desempeño de los estudiantes durante el curso de estudio, se podrá crear un modelo de predicción mejor y más útil.

El paso de la comprensión de los datos es esencial para descubrir patrones de datos útiles con éxito [25]. Los datos educativos utilizados en este documento provienen de 261 estudiantes únicos inscritos en el curso introductorio de Sistemas de bases de datos en la Universidad Constantine the Philosopher en Nitra entre los años 2016 y 2020. Los datos se anonimizaron durante el proceso de extracción.

Posteriormente, se determinó la proporción de estudiantes desértos en cada año académico. Como resultado, el 21,92% del total de estudiantes no logró aprobar el curso, como se muestra en la Figura 1.

Posteriormente, se determinó la proporción de estudiantes desértos en cada año académico. Como resultado, el 21,92% del total de estudiantes no logró aprobar el curso, como se muestra en la Figura 1.

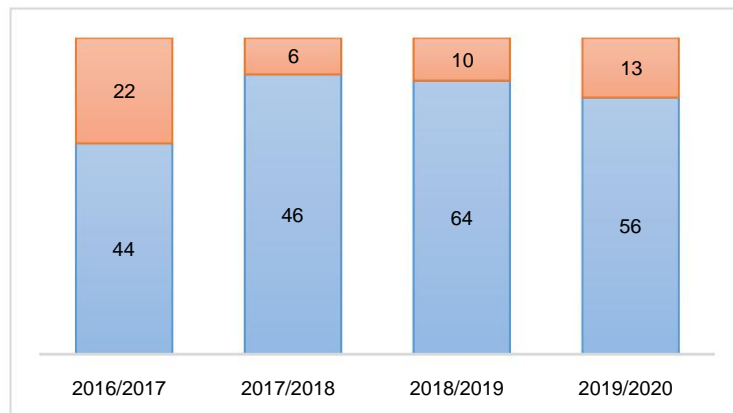


Figura 1. Distribución de la deserción de los estudiantes en varios años académicos (naranja—número de reprobados estudiantes, azul: número de estudiantes que terminaron con éxito el curso).

Las razones de la deserción de los estudiantes fueron diferentes. El número insuficiente de puntos obtenidos, trabajo retrasado no entregado, así como baja actividad en el curso representaron las razones más frecuentes. Sin embargo, una razón particular de abandono no ha sido la más analizada y pena de ser que nadie en la investigación en curso.

Una de las preguntas abiertas de la ERM ya mencionadas es cómo predecir el rendimiento de los estudiantes en función de su actividad y los resultados parciales durante el estudio. El desempeño de los estudiantes se evalúa con los atributos disponibles en el curso. El desempeño ideal es que identifiquen las afirmaciones finales como las afirmaciones correspondientes al objetivo era dar una explicación de por qué se puede explicar el resultado mediante atributos dados la finalización fallida del curso.

El curso se dividió en 40 secciones de acuerdo con los estándares universitarios. Cada una de las secciones contenía tareas, así como una o más de tres tareas en forma de diferentes actividades y cuestionarios de LMS Moodle de las asignaciones de LMS Moodle. Las asignaciones se utilizaron predominantemente en el primer mes del período de seguimiento, seguidas por la prueba parcial y la asignación del proyecto.

3.3. Limpieza de datos y selección de características

La preparación de datos implica la limpieza de datos, la unión de datos, la transformación de atributos y la selección o reducción de características. Se pueden aplicar varios métodos para analizar los datos basados en la comprensión de los datos para descubrir patrones ocultos [27]. El resultado de esta fase es una construcción del conjunto de datos final que se considera importado de los datos sin procesar iniciales. La fase es una construcción del conjunto de datos final que se considera importado desde el bruto inicial a la herramienta de modelado.

Los datos educativos a menudo son escasos, ruidosos, inconsistentes o incluyen varios irrelevantes. Por lo tanto, este paso suele ser aquel donde el objetivo es volver a eliminar todas las características de datos no deseadas. Los datos ruidosos se pueden suavizar agrupando,

regresión o agrupamiento. También hay varias técnicas sobre cómo superar la

problema de los valores perdidos. Los registros incompletos pueden ser removidos o imputados con un constante global o el valor más probable [9]. Sin embargo, la semántica de la falta

valores deben tenerse en cuenta en todos los casos de forma individual.

Como ya se mencionó, los datos educativos utilizados en este trabajo representan 261 estudiantes matriculados en el curso de aprendizaje electrónico utilizado en forma de estudio combinado durante cuatro académicos años. Los resultados de las actividades de aprendizaje interactivo se exportaron utilizando un herramienta de LMS Moodle. Por lo tanto, no fue necesario identificar a los estudiantes individuales. El

los datos personales han sido eliminados.

Simultáneamente, los registros detallados sobre su actividad en forma de solicitud También se exportaron registros. Sin embargo, estos datos tuvieron que ser limpiados en contraste con los grados recibida por los alumnos. La razón es que estos registros contenían información sobre otras partes interesadas como maestros y administradores, así como información sobre actividades,

que no se relacionan directamente con el proceso de aprendizaje. Después de la limpieza, los registros se agregaron para cada estudiante y se unieron con sus resultados parciales.

El número de registros representa un problema persistente en el campo de investigación de tecnología educativa, análisis de aprendizaje y minería de datos educativos [18]. Este problema se puede superar si los métodos de investigación utilizan únicamente los registros, que representan las huellas digitales del comportamiento de las partes interesadas del EVA. Por otro lado, si el conjunto de datos de entrada también debe contener los logros o calificaciones de los estudiantes, la cantidad de registros únicos se reduce radicalmente y los registros deben agregarse [29]. La razón es que los datos sobre los logros parciales de los estudiantes a menudo deben agregarse para minimizar el problema de datos incompletos, la presencia de valores faltantes o el carácter voluntario de las tareas que el estudiante podría entregar. Los valores que faltan no se pueden eliminar fácilmente debido a su semántica. Es posible que no se supiera si los estudiantes entregaron la tarea y recibieron cero puntos, no participaron en la actividad o seleccionaron otra tarea voluntaria. Al mismo tiempo, este registro no se puede eliminar del tamaño ya limitado del conjunto de datos.

Teniendo en cuenta las revisiones de investigación en el dominio de análisis de aprendizaje y minería de datos educativos, el conjunto de datos utilizado en este estudio de caso puede considerarse suficiente [18]. Al mismo tiempo, es necesario tener en cuenta la reducción de la interpretación de los resultados y las opciones más limitadas para generalizar los resultados obtenidos cuando se aplican técnicas de aprendizaje automático para conjuntos de datos tan escasos [30].

Los modelos de predicción precisos y robustos dependen de características minuciosamente seleccionadas extraídas del conjunto de datos estudiado [8]. Por lo tanto, el proceso de extracción de características es fundamental, ya que afecta directamente el rendimiento de cualquier modelo de predicción. De acuerdo con trabajos previos, diferentes estudios a menudo usan diferentes características [7]. Sin embargo, todos generalmente se refieren a las interacciones de los estudiantes con el contenido del curso y los rastros de sus actividades en la plataforma. Representan las características de comportamiento natural que la plataforma LMS Moodle realiza un seguimiento [31,32].

Como se mencionó anteriormente, el curso utilizado como fuente principal de datos contenía 13 secciones con más de 30 actividades interactivas, que requerían la actividad continua de los estudiantes. Estas actividades se pueden dividir en tareas, pruebas, proyecto y examen. Los resultados parciales obtenidos de estas actividades y el número de accesos al curso podrían considerarse a primera vista un conjunto ideal de características de entrada para los algoritmos de aprendizaje.

Por lo tanto, estaban destinados a procesarse mediante la selección de características estándar o la técnica de reducción de la dimensionalidad. En consecuencia, se consideraron diferentes niveles de granularidad de características, desde los niveles de actividad individuales, hasta diferentes tipos de su categorización o agregación con el fin de aplicar alguna técnica de selección de características.

Aunque la importancia de esta técnica aumenta con el número de características en el conjunto de datos de entrada, el número restringido de características y registros influyó en la decisión final de no utilizar la técnica de selección automática de características y reemplazarla con la agregación basada en la categoría de actividad. En otras palabras, si bien los puntos parciales obtenidos de las actividades individuales podían ser considerados en el conjunto de datos de entrada como características individuales, fueron reemplazados por la suma total de la categoría de actividad en el período observado del término. Como resultado, la aplicación del método de selección de características automatizadas no fue necesaria para este estudio debido a un número limitado de variables independientes, que se crearon a partir de las características originales en función de su agregación.

Además de la cantidad de funciones, otra razón de esta decisión fue que los puntos tenían el mismo impacto en la calificación final. Las tareas y las pruebas se distribuyeron uniformemente en las primeras ocho secciones del curso, y su secuencia no jugó un papel significativo en esta investigación.

La razón final es que, como se mencionó anteriormente, el curso se usó durante cuatro años académicos. Durante este período, no siempre se utilizaron todas las tareas y pruebas. Como resultado, varias características deben eliminarse o requerir la aplicación de métodos de imputación adecuados para que se incluyan correctamente en el conjunto de datos. Estos enfoques darían como resultado una mayor reducción de un conjunto de datos ya limitado. Por lo tanto, los resultados parciales de diferentes actividades

fueron agregados, lo que refleja que las asignaciones parciales tenían la misma parte de la final nota en los cursos en todos los años académicos.

Simultáneamente, técnicas alternativas de agregación de características de entrada basadas en el curso Se examinaron temas y semanas durante el semestre. Sin embargo, finalmente se consideraron menos perspectivas que la agregación basada en las categorías.

Como resultado, los siguientes cinco atributos que reflejan las categorías de las actividades del curso fueron considerados en la investigación presentada (Tabla 1):

- Los accesos representan el número total de visitas al curso por parte de un estudiante en el periodo.
- Las tareas representan una puntuación total de diferentes tipos de actividades evaluadas dentro el periodo observado.
- Las pruebas representan una puntuación total de las pruebas parciales y finales durante el semestre.
- También se consideraron las características del examen y del proyecto. Si bien ofrece un alto valor analítico, éstos no se incluyeron por haberse realizado en el último tercio del plazo, y su Los valores están directamente relacionados con completar el curso. Esos valores no estaban disponibles en el momento de la predicción esperada (durante un semestre cuando las funciones Proyecto y Examen están desconocido todavía).

Tabla 1. Estadísticas descriptivas de variables independientes para el conjunto de datos de los años 2016–2019.

	Acceso	Pruebas	asignaciones	Examen	Proyecto
contar	261.000	261.000	216.000	261.000	261.000
media	680.295	37.941	18.077	11.555	65.513
std	374.662	13.113	10.831	6.303	30.615
min	13.000	0.000	0.000	0.000	0.000
25%	440.000	34.079	11.000	9.000	66.410
50%	605.000	40.500	14.400	14.000	75.140
70%	898.750	46.500	24.000	16.000	84.610
máximo	2392.000	57.000	40.000	20.000	99.480

Se realizó una correlación estándar de Pearson para encontrar las correlaciones entre los variables independientes individuales y resultados finales, representados por la variable `result_points`.

Esta variable representa la suma total de puntos parciales, que el estudiante podría obtener durante el curso. Se presenta una matriz de correlación, creada utilizando la biblioteca Seaborn en Python. en la Figura 2. Las siguientes conclusiones se pueden extraer de los valores de correlación representados en la figura 2.

- existe una correlación media entre los atributos clave, con los valores correlacionados con el número resultante de puntos al final del curso. Este hallazgo no es sorprendente,
- las variables proyecto y pruebas tienen una correlación ligeramente más fuerte con `result_points` que los otros atributos. Puede deducirse que tienen un impacto más crítico en la resultado final del estudiante. Por lo tanto, deben ser considerados a fondo con respecto a su ocurrencia en el curso. Si se utilizan más adelante en el curso, tendrían una mayor impacto significativo en la precisión de la predicción.

Debido a que la predicción debe hacerse lo antes posible para tener suficiente tiempo para intervención, las variables acceso, asignaciones y pruebas fueron finalmente seleccionadas como las dos características más importantes. Es necesario recalcar, que sólo suma de pruebas parciales, que se realizaron en el período observado, se tuvo en cuenta. También se puede afirmar que un atributo más apropiado para la predicción, además de los dos atributos mencionados con un alto correlación (proyecto y pruebas), es también el número de accesos (acceso). Un posible uso de este atributo puede ser para identificar a los estudiantes que no alcanzan el tiempo mínimo permitido (presencia o frecuencia de accesos), exponiéndolos así a un mayor riesgo de obtener una Evaluación del punto más débil.

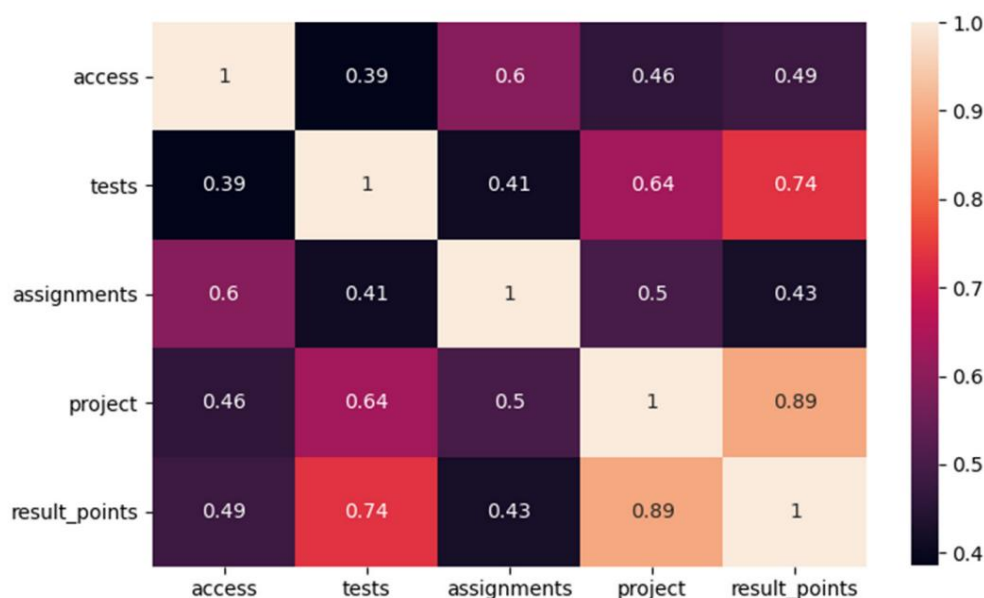


Figura 2. Una matriz de correlación entre variables de entrada independientes.

Figura 2. Una matriz de correlación entre variables de entrada independientes.

Un modelo de aprendizaje automático supervisado también requiere un ejemplo de resultados (objetivo) de la que puede aprender además de los datos de entrada. Dado que la información sobre la final

Debido a que la predicción debe hacerse lo antes posible para tener suficientes calificaciones disponibles de los estudiantes, fue posible crear una variable objetivo binaria. Cada

intervención, finalmente, se seleccionaron las variables de acceso a áreas de pruebas y a que al estudiante cuya calificación final fue de 15 (suspensión) se le asignó un valor numérico de 0, mientras que

aprobatorias se asignó un valor de 1. Como se mencionó anteriormente, las calificaciones binarias, a los alumnos

fue leerlos en el período observado, se tuvo en cuenta. También se puede afirmar t Los datos utilizados en el

estudio de caso constaban de dos partes. La primera parte, utilizada como entrenamiento

El atributo más apropiado para la predicción, además de los dos atributos mencionados y el conjunto de datos de prueba, contenía datos de los tres años académicos (2016-2019). El segundo

alta correlación (*proyecto y pruebas*), es también el número de accesos (*acceso*). Una posible parte se formó a partir de la ejecución más reciente del curso en el año 2020. Estos datos fueron

de este atributo puede ser identificar a los estudiantes que no alcanzan el tiempo mínimo utilizado para evaluar el desempeño de los modelos de clasificación final (Tabla 2).

(presencia o frecuencia de accesos), exponiéndolos así a un mayor riesgo de obta **Tabla 2.** Estadísticos descriptivos de variables independientes para el conjunto de datos del año 2020.

	Un modelo de aprendizaje automático supervisado que también requiere un ejemplo de los datos de salida.	Tareas de
Dado que la información sobre las calificaciones de los estudiantes estaba disponible, fue posible crear una variable objetivo binaria.	media 925.533 std 309.606 valor mínimo 694.250 valor máximo 1152.500	estudiantes 45.114 disponible 25.533 se le asignó un valor numérico de 1 35.00 se le asignó un valor numérico de 0 8.166
al estudiante cuya calificación final fue "FX" (reprobando) se le asignó un valor numérico de 0,00, los estudiantes exitosos se le asignó un valor numérico de 1. De esta manera se logró el clasificado binario 25% 50% 1 70%.	74.405 33.238 76.800 88.470 94.555	estudiantes 11.842 disponible 5.192 se le asignó un valor numérico de 1 1.334 se le asignó un valor numérico de 0 34.00
Los datos utilizados en el estudio de caso constaban de dos partes. La primera parte, utilizada como un conjunto de entrenamiento, consistió en 80% de los datos.	74.405 33.238 76.800 88.470 94.555	estudiantes 11.842 disponible 5.192 se le asignó un valor numérico de 1 1.334 se le asignó un valor numérico de 0 34.00

de datos de tra y prueba, contenía datos de los tres años académicos (2016-2019). El conjunto de datos de evaluación

parte se formó a partir de la ejecución más reciente del curso en el año 2020. Este dato en general se comparó con los datos de la Tabla 1. Estos cambios en un número absoluto de puntos utilizado para evaluar el desempeño de los modelos de clasificación final (Tabla 2), se establecieron en el último tramo del curso con el objetivo de practicar las habilidades de los alumnos.

El conjunto de datos de evaluación (Tabla 2) mostró un aumento sustancial en la media y los valores medios de la categoría proyecto y pruebas aumentaron un 14 % y un 18 %, respectivamente. En general, el acceso y asignaciones lograron valores medios considerablemente más altos del 56 % estándar antes del inicio que para la evaluación de la propuesta. practicar los alumnos y el 42%. Tuvieron que ser modelos de clasificación.

Los valores medios de la categoría proyecto y pruebas aumentaron un 14% y un 16%, respecto a este hecho. El paso final de la preparación de datos fue la normalización y estandarización. Las categorías de asignaciones lograron valores medios considerablemente más altos o la representación de requisitos comunes de muchas máquinas de aprendizaje y 42%. Tuvieron que ser estandarizados antes de su uso para la evaluación de los estimadores pro. Si la característica específica no se parece a los datos estándar normalmente distribuidos, modelos de clasificación, los

Estimadores pueden comportarse incorrectamente [33]. Los datos se transformaron usando Standard Escalador para estandarizar características eliminando la media y escalando a la varianza de la unidad. Centrado y la escala se produjo de forma independiente en cada función mediante el cálculo de las estadísticas relevantes en el conjunto de datos.

4. modelo

La fase de modelado consiste en aplicar diferentes técnicas de aprendizaje automático al conjunto de datos. El objetivo de la predicción es crear un modelo basado en las actividades y logros actuales de los estudiantes que intente predecir el fracaso del estudiante y el desempeño futuro. Es un problema de clasificación típico, que un modelo de clasificación binaria o un clasificador binario puede resolver para predecir si un estudiante puede completar el curso o no.

Hoy en día, existen muchos métodos diferentes de clasificación de aprendizaje automático, que se pueden utilizar para encontrar un modelo predictivo adecuado [3]. Teniendo en cuenta los requisitos resultantes del objetivo principal del artículo y trabajos relacionados [28], finalmente se aplicaron los siguientes clasificadores más adoptados: 1. La regresión logística (LR) es un algoritmo de clasificación también conocido como regresión logit, que puede calcular una probabilidad predicha para una variable dependiente dicotómica basada en una o más variables independientes [7,34].

2. Un árbol de decisión (DT) se define dividiendo recursivamente el espacio de entrada y definiendo un modelo local en cada región resultante del espacio de entrada. Este algoritmo divide esta tarea de clasificación compleja en un conjunto de tareas de clasificación simples [31].
3. El clasificador Naïve Bayes (NB) es un clasificador probabilístico simple con características estrictas suposiciones de independencia basadas en el teorema de Bayes [4,35].
4. La máquina de vectores de soporte (SVM) se diseñó inicialmente para la clasificación binaria mediante la construcción o el establecimiento de un hiperplano ideal que divide el conjunto de datos en clases [34].
5. Random Forest (RF) es una técnica que utiliza múltiples árboles de decisión diseñados para realizar la estimación utilizando bootstrapping, remuestreo y luego aplicando la votación mayoritaria o promediando [2,31].
6. Una red neuronal (NN) es un sistema de procesamiento paralelo que aprovecha la estructura y las funciones del cerebro para maximizar la precisión predictiva. El aprendizaje profundo NN toma datos como entrada y luego los procesa para calcular la salida en varias capas [31,33].

La idea principal es encontrar el modelo de predicción más adecuado y comparar las diversas métricas de rendimiento de predicción de cada clasificador aplicado. Los hiperparámetros se ajustaron utilizando la técnica de búsqueda en cuadrícula, que se usa ampliamente para encontrar la mejor configuración de parámetros.

Se utilizó el lenguaje de programación Python. Python ofrece bibliotecas ricas y eficientes para aplicar métodos de aprendizaje automático como las bibliotecas Pandas y Numpy para la manipulación de datos. Posteriormente, se utilizó la biblioteca Scikit-learn que contiene varios conjuntos de métodos de aprendizaje automático supervisados o no supervisados. Se utilizaron otras bibliotecas, Keras y Tensorflow, para construir una red neuronal. Keras es una API para Tensorflow, lo que facilita la creación de NN [32,36].

Hoy en día, hay disponibles varias bibliotecas de aprendizaje automático de código bajo que automatizan el flujo de trabajo del aprendizaje automático. Su aplicación en el dominio LA no es frecuente. Por lo tanto, considerando el objetivo principal del artículo, que discute las limitaciones de la aplicación de los clasificadores en el escaso conjunto de datos educativos, se agregaron varios clasificadores utilizando la biblioteca PyCaret [37].

Es un desafío crear un punto de predicción preciso adecuado para identificar a los estudiantes en riesgo de abandonar un curso si el conjunto de datos es más pequeño. Los métodos de clasificación mencionados anteriormente se compararon usando datos divididos en conjuntos de entrenamiento y prueba. Los conjuntos de prueba y entrenamiento se eligieron al azar para obtener intervalos de confianza para todas las siguientes métricas de rendimiento [34]: La precisión es la tasa de precisión general o la precisión de la clasificación, y se determina de la siguiente manera:

$$\text{Precisión} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

La tasa de verdaderos positivos (tasa de TP) o recuperación es la proporción de positivos reales que se predice que serán positivos. En este estudio de caso, el retiro se utilizó para evaluar a los que completaron. Se determina de la siguiente manera:

$$\text{recordar} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

La tasa negativa verdadera (tasa TN) o precisión es la proporción de negativos reales que se espera que sean negativos. En este trabajo se utilizó la precisión para evaluar la deserción de los estudiantes, y es determinado de la siguiente manera:

$$\text{Precisión} = \frac{\text{Tennesse}}{\text{TN} + \text{FP}}$$

F1score muestra el equilibrio entre dos medidas de clasificación. Representa una medida de cálculo de compensación ampliamente utilizada para conjuntos de datos desequilibrados, y se determina de la siguiente manera:

$$\text{Puntuación F1} = \frac{2 \times \text{Recordar} \times \text{Precisión}}{\text{Recordar} + \text{Precisión}}$$

Finalmente, la tasa de error de clasificación representa una proporción de instancias mal clasificadas sobre todo el conjunto de instancias. Se calcula de la siguiente manera:

$$y = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Este estudio debe considerar varias métricas para estimar el desempeño de los diferentes métodos para clasificar correctamente los casos en un conjunto de datos independiente y evitar el sobreajuste.

Todos los algoritmos de clasificación se ejecutaron utilizando un procedimiento de validación cruzada de 10 veces. El proceso de validación cruzada de 10 veces divide la recopilación de datos en diez secciones aproximadamente iguales. Las nueve partes restantes se utilizan para entrenar el modelo y el error de prueba se calcula clasificando la parte dada. Por último, se promedian los resultados de diez pruebas.

5. Resultados y Evaluación

Los modelos producidos en la fase anterior deben evaluarse para una implementación satisfactoria en esta etapa de la metodología CRISP-DM. La evaluación utilizada en este caso de estudio fue numérica basada en métricas de rendimiento del modelo, en contraste con la evaluación empírica, en la que el creador del modelo y el experto en conocimiento del dominio evalúan si el modelo se ajusta al propósito.

Se utilizaron los datos más recientes del curso del año académico 2020 para evaluar el desempeño de cada algoritmo de clasificación. Estos datos se denominan más adelante como datos no vistos.

5.1. Comparación de métricas de

rendimiento La Tabla 3 compara los resultados de clasificación de la prueba de validación cruzada de 10 veces obtenidos por Regresión logística (LR), Árbol de decisión (DT), Bosque aleatorio (RF), Naïve Bayes (NB), Máquinas de vectores de soporte (SVM) y Algoritmos de redes neuronales (NN) para el curso seleccionado, que informan sobre algunas de las medidas/indicadores de éxito más populares: exactitud, precisión, recuperación y puntuación F1.

El clasificador Naïve Bayes obtuvo los resultados de mayor precisión en el conjunto de datos de evaluación con una tasa de éxito del 93 %. Sin embargo, el error de clasificación fue de 0,231, con una precisión global del modelo del 77 %. Teniendo en cuenta este hecho, este algoritmo puede haber sufrido por la fuerte suposición de independencia en el conjunto de datos e identificó muchos casos de falsos negativos, ya que el valor de recuperación alcanzó solo el 72 %, lo que redujo la precisión general. La puntuación de F1 alcanzó el 82%. Por lo tanto, no se sugiere como un clasificador adecuado para predecir la deserción de los estudiantes en este caso.

Tabla 3. Comparación de las medidas de rendimiento de los algoritmos de clasificación aplicados.

Modelo	Exactitud	Recuerdo	Precisión	Puntuación F1
Naïve Bayes (NB)	0.77	0.72	0.93	0.82
Bosque aleatorio (RF)	0.93	0.96	0.86	0.91
Red neuronal (NN)	0.88	0.89	0.86	0.88
Regresión Logística (LR)	0.93	0.98	0.79	0.90
Máquinas de vectores de soporte (SVM)	0.92	0.96	0.79	0.88
Árbol de decisión (DT)	0.90	0.98	0.71	0.85

De acuerdo con la Tabla 3, el clasificador de bosque aleatorio obtuvo la mejor exactitud, precisión, Puntuaciones F1 y el segundo mejor valor de recuperación. Vale la pena mencionar que RF gana una ventaja sobre otros clasificadores gracias a varios hiperparámetros que se pueden ajustar y conducir a predicción más precisa. Los hiperparámetros resultantes de la búsqueda en cuadrícula señalaron la construcción de 1000 árboles dentro del modelo de bosque con una profundidad máxima de 50 por cada árbol. Gracias a estos ajustes, la precisión del modelo alcanzó el 86%, la recuperación fue del 96 %, mientras que la puntuación F1 fue del 91 %. El error de clasificación fue de 0,066, con un total precisión del 93%.

El modelo de red neuronal construido en Keras se describe como una serie de capas. El La función de activación de ReLU se usó en ambas capas ocultas. Debido a la clasificación binaria problema, la función de activación sigmoide se utilizó en la capa de salida para garantizar la salida valores entre 0 y 1. Se ha elegido el optimizador “adam” ya que ajusta automáticamente mismo y dio buenos resultados en una amplia gama de problemas. Teniendo en cuenta que el conjunto de datos es relativamente pequeño, y el proceso de entrenamiento pasará por un número fijo de iteraciones, el número 50 se seleccionó con un tamaño de lote de 10, que es el número de filas del conjunto de datos antes de cambiar los pesos del modelo para este problema. Con todos esos ajustes, el error de clasificación fue de 0,122, con una precisión global del 88%. la red neuronal logró una puntuación F1 del 88%. La recuperación alcanzó el 89 %, mientras que la precisión finalizó con un 86 % de exactitud.

No hay hiperparámetros críticos para ajustar dentro del clasificador de regresión logística. No se lograron diferencias útiles en el rendimiento con varios solucionadores, regularización, y el parámetro C, que controla la fuerza de regularización. La precisión general aumentó en un 2-3%, pero el valor de recuperación disminuyó significativamente durante la optimización de hiperparámetros. Por lo tanto, en este caso, los resultados finales logrados por LR se obtuvieron bajo el valor predeterminado escenarios donde las puntuaciones de precisión y recuerdo fueron 79% y 98%, respectivamente. El el error de clasificación fue de 0,067, con una precisión global del 93 %.

El algoritmo de clasificación de máquinas de vectores de soporte es muy común y tiene varios hiperparámetros para ajustar el resultado. La elección del núcleo, que controlará cómo se Se proyectan las variables de entrada, es quizás el parámetro más importante. El parámetro C También debe considerarse como importante elegir. Representa la pena y puede contener una variedad de valores. Tiene un impacto significativo en la forma de las regiones resultantes para cada clase. En este caso, los hiperparámetros resultantes obtenidos por la búsqueda en cuadrícula apuntaron a kernel lineal con un valor de penalización de 1.0. El error de clasificación fue de 0.073, con un promedio precisión del 92%. El algoritmo logró una puntuación F1 del 88 %, mientras que la recuperación y la precisión las puntuaciones alcanzaron el 96% y el 79%, respectivamente.

El tamaño del árbol, también conocido como el número de capas, es el árbol de decisión más crítico. parámetro del modelo. Se espera que los árboles poco profundos, también conocidos como aprendices débiles, se desempeñen mal porque no capturan suficiente información sobre el problema. En el otro Por otro lado, los árboles más profundos capturan demasiados aspectos, lo que lleva a sobreajustar el conjunto de datos de entrenamiento, lo que dificulta hacer una predicción precisa sobre datos no vistos. Como resultado, la decisión modelo de árbol fue creado bajo el parámetro de profundidad de 4. Este modelo obtuvo 90% en general la precisión y el error de clasificación fue de 0,103. Si bien el recuerdo alcanzó un valor prometedor del 98%, la precisión alcanzó solo el 71 %, lo que también afectó la puntuación más baja de F1 del 85 %.

5.2. Comparación de clasificadores de aprendizaje automático usados

El rendimiento de los clasificadores seleccionados también se puede comparar utilizando estadísticas pruebas de importancia para comparar algoritmos de aprendizaje automático, aunque el conjunto de datos usado en este estudio de caso fue pequeño y el entrenamiento de múltiples copias de modelos clasificadores no fue costoso o poco práctico. Como cada modelo se evaluó en el mismo conjunto de datos, McNemar's La prueba podría usarse para comparar los algoritmos de clasificación si los modelos parecen predecir de la misma manera o no. La prueba de McNemar es una prueba no paramétrica pareada o sin distribución. prueba de hipótesis estadística [38]. Esta prueba comprueba si los desacuerdos entre dos casos coincidir. Por lo tanto, la prueba de McNemar es un tipo de prueba de homogeneidad para tablas de contingencia.

Los resultados de la prueba de McNemar (Tabla 4) muestran que los clasificadores emparejados con un valor p superiores a 0,5 habían infravalorado y sobrevalorado sistemáticamente el mismo conjunto de pruebas. Eso significa que los clasificadores tienen una proporción de error similar en el conjunto de prueba. Simultáneamente, el los resultados muestran que no son significativas en la mayoría de los casos.

Tabla 4. Comparación de la prueba de McNemar de los clasificadores.

Modelo	LR	DT	RF	naïve bayes	MVS
LR	-	p = 1.000	p = 0,500	p = 0,001	p = 1.000
DT	-	-	p = 0,250	p = 0,000	p = 1.000
RF	-	-	-	p = 0,007	p = 0,250
naïve bayes	-	-	-	-	p = 0,000
MVS	-	-	-	-	-

Por otro lado, todos los clasificadores emparejados con el modelo NB habían obtenido un valor p inferior a 0,5. Si diferentes modelos cometen diferentes errores, un modelo es más preciso que el otro y lo más probable es que tenga una proporción diferente de errores en el conjunto de prueba. En ese caso, el resultado de la prueba es significativo. Por lo tanto, como ya se supuso, el clasificador NB es considerado no adecuado para predecir la deserción de los estudiantes en este caso.

6. Discusión

Se compararon seis de los algoritmos de aprendizaje automático más conocidos utilizando un conjunto de datos sobre la actividad de 261 estudiantes de tercer año del curso universitario. El conjunto de datos fue preprocesado de forma estándar. Los modelos fueron entrenados y probados en el conjunto de datos de tres años académicos. Los resultados obtenidos se basan en un conjunto de datos equilibrado. Medidas de éxito se han registrado en ambas categorías, los que completaron y los que no completaron. Ambos se evitaron los sesgos en términos de un conjunto de datos desequilibrado, así como en base al promedio. Posteriormente, se evaluaron los modelos con datos no vistos del cuarto año académico. El Los clasificadores se evaluaron utilizando varias métricas de rendimiento y utilizando la prueba de McNemar, que comparó su capacidad para clasificar los casos no vistos en las mismas clases.

La mejor predicción de deserción basada en la comparación de la precisión, recuperación, exactitud y las puntuaciones F1 se obtuvieron mediante el clasificador de bosque aleatorio (RF) en general. Los resultados obtenidos por los algoritmos LR, SVM y DT fueron los segundos mejores con ligeras diferencias en comparación a RF.

Por otro lado, el Naïve Bayes y el modelo Neural Network tuvieron los peores resultados en general y no se consideran modelos de clasificación precisos para este conjunto de datos limitado. Este hallazgo confirma que el rendimiento de los modelos de aprendizaje automático supervisado depende estrechamente de suficientes registros. Los datos de la evaluación fueron ligeramente diferentes de los datos de entrenamiento porque el curso ha recibido cambios menores en comparación con las ejecuciones anteriores. La precisión de la NN se vio afectada principalmente por estos cambios, ya que el conjunto de datos consta de relativamente pocos registros. Se consideró la estandarización o normalización del conjunto de datos para mejorar el modelo NN. Finalmente, se optó por la estandarización de atributos por ser recomendable y útil para modelos que dependen de la distribución de atributos.

La precisión de la predicción varió entre 77 y 93%, lo que indica que independientemente de el modelo utilizado, las características elegidas en este estudio demostraron ser exitosas para predecir el éxito o la deserción de los estudiantes a pesar de la limitación del conjunto de datos y el número de características

Por otro lado, todos los algoritmos lograron valores de recuperación superiores a 0,77 y simultáneamente una muy buena precisión. Sin embargo, los algoritmos de clasificación podrían lograr una precisión muy alta al predecir solo la clase mayoritaria. Por ejemplo, si alguno de los algoritmos utilizados predice que todos los estudiantes serán "completadores '10", se espera que la precisión del conjunto de datos sea del 78,5 % porque el 21,5 % de los estudiantes abandonaron debido a la distribución estadística.

Los modelos predictivos examinados fueron capaces de alcanzar un alto rendimiento en cada clase (completadores '10 y no completos '00) considerando los valores de recuperación y precisión, como se muestra en la Tabla 3. La precisión actual de las predicciones de los algoritmos de clasificación en esta etapa se debe confiar

En resumen, existe una convergencia entre las métricas de los modelos RF, LR, SVM y DT. Además, la media armónica de precisión y recuperación estuvo representada por la puntuación F. Además, la puntuación F representó la media armónica de precisión y recuperación, lo que indica la tasa general de éxito de los modelos de predicción para predecir los estudiantes que reprobaron. Teniendo en cuenta los valores de puntuación F que se muestran en la Tabla 3, los modelos RF, LR, NN y SVM fueron comparativamente útiles para predecir la probabilidad de que los estudiantes abandonaran los estudios.

Se midió el rendimiento de los algoritmos de clasificación propuestos para un análisis más detallado de la eficacia de varios modelos propuestos. Se consideró el área bajo la curva característica operativa del receptor (AUC-ROC) para evaluar el impacto de un conjunto de datos desequilibrado. El AUC-ROC se usa típicamente para abordar datos desequilibrados porque demuestra el equilibrio entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) en un espacio bidimensional [23].

La curva ROC se obtiene representando los valores de los distintos umbrales calculados. El TPR representa el porcentaje de deserción de los estudiantes que el modelo predijo como estados de deserción en esta investigación. Viceversa, la FPR es el porcentaje de estudiantes que no desertaron, pero el clasificador los predijo como desertores. Esta métrica resume la bondad de ajuste de un clasificador [30].

La Figura 3 muestra que el bosque aleatorio (RF) (AUC = 0.968) y la regresión logística (LR) (AUC = 0,924) los clasificadores son significativamente mejores cuando se miden mediante la curva AUC-ROC.

Árbol de decisión (DT) obtenido (AUC = 0,883), seguido de una máquina de vectores de soporte (SVM) (AUC = 0,850) y Naïve Bayes (NB) (AUC = 0,765).

Teniendo en cuenta la distribución de clase positiva (completadores '10) y clase negativa (no completadores '00) en el conjunto de datos investigado, cada uno de los algoritmos de clasificación utilizados es muy bueno para capturar ocurrencias de clase positiva. Sin embargo, algunos de ellos, como DT, SVM y NB, generan muchas predicciones falsas positivas, lo que reduce la precisión del rendimiento predictivo final. El clasificador de bosque aleatorio y la regresión logística pueden distinguir pequeñas diferencias entre los parámetros de acondicionamiento que resultan en menos fallas predichas incorrectamente. Por lo tanto, su rendimiento predictivo es mayor.

Esta declaración también puede confirmar la tasa de error de clasificación, que representa una proporción de casos mal clasificados sobre el conjunto total de casos. Nuevamente, los clasificadores RF y LR lograron los mejores valores en esta métrica de rendimiento. Clasificaron mal solo alrededor del 6% de los casos, lo que contrasta con el clasificador NB, que clasificó mal el 23% de los casos.

Árbol de decisión (DT) obtenido (AUC = 0,883), seguido de una máquina de vectores de soporte (SVM) (AUC = 0,850) y Naïve Bayes (NB) (AUC = 0,765).

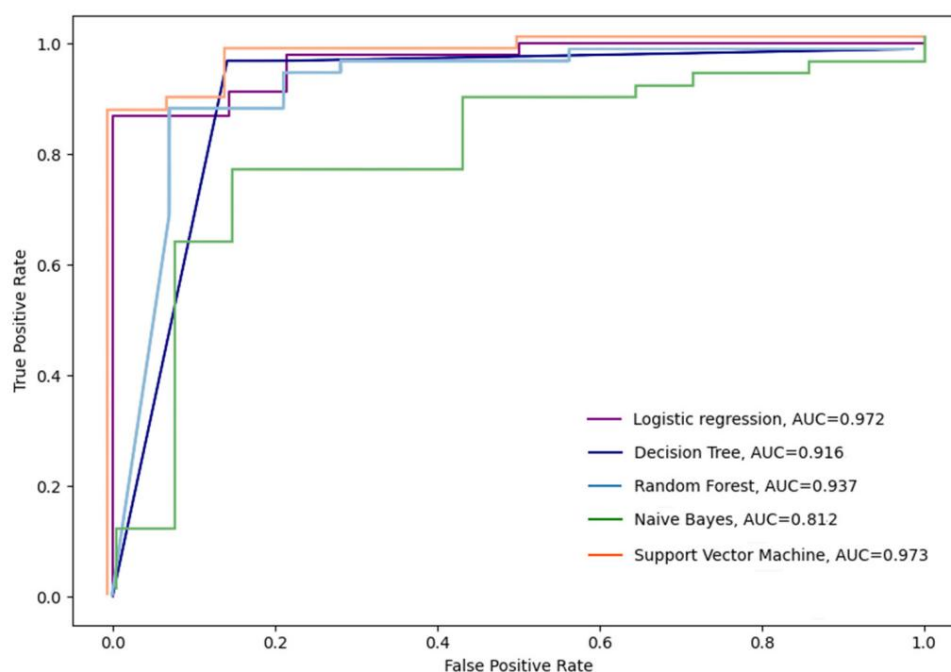


Figura 3. Comparación de las medidas de rendimiento por área bajo las curvas de la curva característica operativa del receptor (AUC-ROC).

La prueba de McNemar proporciona otro punto de vista interesante sobre los clasificadores individuales. La comparación mutua de los clasificadores binarios que se muestra en la Tabla 4 descubrió que el RF. Teniendo en cuenta la distribución de clase positiva (completadores '1') y clase negativa (los clasificadores no y LR clasifican los casos individuales no vistos en las mismas clases. No hay completadores '0') en el conjunto de datos investigado, cada uno de los algoritmos de clasificación utilizados, es una diferencia muy significativa entre ellos. Simultáneamente, no hay diferencia significativa bueno para capturar ocurrencias de clase positivas. Sin embargo, algunos de ellos, como DT, SVM, entre otros clasificadores, a saber, entre DT y SVM, DT y LR, y entre LR y NB, generan muchas predicciones de falsos positivos, lo que reduce el rendimiento predictivo final y la SVM. precisión de mance. El clasificador de bosque aleatorio y la regresión logística pueden distinguir pocas diferencias. Finalmente, se comparó el costo computacional de ejecutar todos los algoritmos de investigación. referencias entre los parámetros condicionantes que dan como resultado menos predicciones incorrectas en aras de la exhaustividad. Los cuatro algoritmos de clasificación se ejecutaron en menos de 1,00 s, fallas Por lo tanto, su rendimiento predictivo es mayor, como se muestra en la Tabla 5. Naturalmente, el NN registró un tiempo significativamente mayor debido a Esta declaración también puede confirmar la tasa de error de clasificación, que representa una proporción de casos mal clasificados sobre el conjunto total de casos. Nuevamente, los clasificadores RF y LR lograron los mejores valores en esta métrica de rendimiento. Clasificaron mal solo alrededor del 6% del de los casos, lo que contrasta fuertemente con el clasificador NB, que clasificó erróneamente el 23% de la casos. Tiempo de ejecución (en segundos).

Modelo	tiempo de ejecución
LR	0.36
DT	0.84
SVM	0.38
MVS	0.52
NN	8.95

Finalmente, se comparó el costo computacional de ejecutar todos los algoritmos de investigación para completarlo. Los cuatro algoritmos de clasificación se ejecutaron en menos de 1,00 s, como se muestra en la Tabla 5. Naturalmente, el NN registró un tiempo significativamente mayor debido a las capas ocultas. Sin embargo, dado el marco de tiempo del problema propuesto, su tiempo de ejecución fue no prohibitivo. Tomó alrededor de nueve segundos ejecutarse mientras usaba todos los atributos en el nivel de curso individual, incluso en el caso del escaso conjunto de datos con un número limitado de características de entrada. Sin embargo, es necesario explorar un conjunto más extenso de rendimiento peor de los casos.

Se deben implementar métricas antes de los clasificadores para la predicción de la deserción de los estudiantes. Esta declaración está en línea con los resultados de otros trabajos de investigación publicados en el aprendizaje análisis y dominio de minería de datos educativos.

Sin embargo, el estudio de caso presentado y sus hallazgos tienen algunas limitaciones. Como ya se mencionó, el tamaño limitado del conjunto de datos es el primero. En contraste con muchos otros dominios de aplicación de modelos de aprendizaje automático, la cantidad de datos en el dominio educativo no puede incrementarse fácilmente combinando diferentes recursos. El

La razón es que los registros individuales a menudo deben representar los resultados de aprendizaje o el comportamiento de los estudiantes individuales. Como consecuencia, los conjuntos de datos suelen ser más pequeños de lo que realmente necesitarían los algoritmos de aprendizaje automático. A menudo, esta deficiencia solo se puede superar aplicando el diseño de investigación similar y preciso utilizado en la investigación de tecnología educativa "clásica", en la que se estima o garantiza una cantidad suficiente de registros antes del experimento mismo.

El mismo requisito también es válido para el número y la calidad de las variables independientes (características). El proceso de aprendizaje a menudo se basa en la repetición de algunas actividades. Si estas actividades están mutuamente conectadas o condicionadas, su inclusión en la técnica de aprendizaje automático debe analizarse exhaustivamente, nuevamente, antes de que comience el proceso de recopilación de datos.

La selección del umbral de tiempo en el que se realiza la predicción también puede mencionarse como una debilidad. Dado que la variable tiempo no se incluyó directamente, fue necesario encontrar hitos en los que el desempeño de los estudiantes no fuera fácilmente predecible. El final del segundo tercio del trimestre se utilizó como compromiso en relación con la distribución natural de las categorías individuales de actividades en las secciones del curso. Sin embargo, se esperarían mejores resultados si las actividades en las secciones del curso se diseñaran conociendo los requisitos de las técnicas de aprendizaje automático aplicadas.

Otra limitación de esta investigación es que las ejecuciones individuales de los cursos ofrecieron diferentes datos que podrían analizarse. Por lo tanto, fue un desafío determinar qué atributos son lo suficientemente relevantes para la predicción del desempeño del estudiante en general. Para este propósito, se utilizó la metodología CRISP-DM para repetir los pasos de investigación individuales y descubrir patrones ocultos y relaciones significativas dentro de un conjunto de datos seleccionado o comprender los datos que tienen el impacto más significativo para resolver el problema de abandono escolar. Simultáneamente, se debe analizar y estandarizar el rango máximo de puntos que podrían alcanzar los estudiantes.

La selección de los clasificadores utilizados en este caso de estudio puede considerarse la siguiente limitación. Las revisiones de la literatura sobre el estado actual de la investigación y las tendencias en LA y EDM brindan muchos ejemplos de técnicas de aprendizaje automático más o menos avanzadas, que se pueden aplicar para la predicción temprana de la deserción escolar de los estudiantes a nivel de curso. Sin embargo, ninguno de ellos ha proporcionado resultados significativamente mejores hasta el momento. Dado que el objetivo principal del artículo no era encontrar el mejor, la selección final del clasificador utilizado en este caso de estudio permite señalar que también se podría lograr una buena predicción utilizando clasificadores más simples, pero se deben evaluar las métricas de rendimiento.

La última debilidad del estudio de caso es una forma de intervención adecuada, que no se discute en detalle. El estudio de caso ya ha demostrado que el fracaso del estudiante se puede predecir en función de las categorías de actividades seleccionadas. Por otro lado, la razón de este estado poco favorecedor permanece abierta y requiere más investigación. Sería interesante examinar si otras categorías de actividades deberían tener un impacto similar en la participación de los estudiantes y por qué categorías de actividades pueden intercambiarse en la fase de intervención.

7. Conclusiones

La tasa de deserción de los estudiantes es uno de los principales indicadores de desempeño del e-learning. Es crucial en el caso del entorno virtual de aprendizaje en la universidad, así como en el caso de los MOOC (Massive Open Online Course). La predicción confiable puede alentar a los estudiantes a estudiar más cuando saben que están en riesgo y planean cuidadosamente su carga de trabajo semestral. La dirección de la escuela puede tener en cuenta esta información al decidir si confirma o rechaza las solicitudes de los estudiantes que apelan por repetir cursos en línea reprobados. Predecir a los estudiantes en riesgo de abandono escolar también les da a los educadores una señal para intervenir cuando los estudiantes están en riesgo y mejorar la participación de los estudiantes en actividades interactivas a tiempo. Es fundamental revelar qué tienen en común los cursos y cómo este conjunto de características puede influir en el comportamiento y el rendimiento de los estudiantes.

La investigación presentada contribuye a solucionar el problema de la deserción de los estudiantes a nivel de curso. Los resultados mostraron que los indicadores adecuadamente seleccionados, que no requieren acceso a los registros del sistema, pueden ser beneficiosos a pesar de un pequeño conjunto de datos si se evalúan diferentes métricas de rendimiento. Los datos recopilados sobre las actividades del entorno de aprendizaje en línea de los estudiantes y sus logros parciales se utilizaron para ingresar los modelos predictivos. Simultáneamente, se ha demostrado que una metodología propuesta es confiable para predecir la finalización del curso cuando hay suficiente tiempo para alentar a los educadores a realizar intervenciones oportunas.

Los resultados descubrieron que los modelos de clasificación podrían considerarse lo suficientemente confiables como para hacer una predicción adecuada de finalización o abandono temprano del curso antes del final de un semestre. De hecho, se obtuvieron buenos resultados de predicción, independientemente del modelo empleado. Las características seleccionadas en este estudio demostraron ser influyentes en la predicción de personas que completaron y no completaron. La precisión de la predicción varió entre 77 y 93%.

En este contexto, es importante señalar que reconocer a los estudiantes en riesgo de deserción es solo el primer paso para abordar la deserción escolar. El siguiente paso es reconocer las necesidades y problemas únicos de cada estudiante que abandonan los estudios y luego introducir ajustes para incluir estrategias efectivas de prevención de la deserción [39]. Esto puede ilustrar las razones reales de la deserción de los estudiantes y desarrollar estrategias que puedan alentar a los estudiantes a completar sus cursos a tiempo con éxito. Los instructores también deben poder concentrarse en las necesidades de los estudiantes para ayudarlos a prevenir la deserción a tiempo.

Al mismo tiempo, también vale la pena recordar que el secreto para evaluar el impacto de estos diferentes tipos de intervenciones es encontrar la más adecuada para cada forma de estudiante con riesgo de abandono. Por lo tanto, la investigación futura debe centrarse en aplicar los modelos propuestos para predecir los estudiantes en riesgo de abandonar la escuela en función de los logros anteriores de estudiantes similares. El conocimiento obtenido de esta investigación puede servir como base para diseñar una recomendación de curso.

Estas pautas, sin embargo, deben tener en cuenta los talentos, las habilidades, las preferencias y las actividades de tiempo libre de los estudiantes en sus horarios. Los métodos propuestos con predictores fácilmente obtenibles son bastante generales y pueden utilizarse en otros cursos.

Contribuciones de los autores: Conceptualización, JK y MD; metodología, JK y MD; validación, JK y MD; análisis formal, JK y MD; recursos, JK y MD; curación de datos, JK y MD; redacción—preparación del borrador original, JK y MD; redacción—revisión y edición, JK y MD; visualización, Janka Kabathova; supervisión, MD; administración de proyectos, MD; adquisición de fondos, MD. Todos los autores han leído y están de acuerdo con la versión publicada del manuscrito.

Financiamiento: Esta investigación fue financiada por la Comisión Europea bajo el Programa ERASMUS+ 2018, KA2, número de subvención: 2018-1-SK01-KA203-046382 "Work-Based Learning in Future IT Profession als Education" y la Agencia Cultural y Educativa de la Ministerio de Educación de la República Eslovaca, número de subvención: KEGA029UKF-4/2018 "Métodos innovadores en la educación de programación en la formación universitaria de profesores y profesionales de TI".

Declaración de la Junta de Revisión Institucional: No corresponde.

Declaración de consentimiento informado: Se obtuvo el consentimiento informado de todos los sujetos involucrados en el estudio.

Declaración de disponibilidad de datos: los datos presentados en este estudio están disponibles abiertamente en <https://github.com/JKabathova/Students-Dropout-Prediction> (consultado el 10 de febrero de 2021).

Conflictos de interés: Los autores declaran no tener ningún conflicto de interés.

Referencias

1. Narayanasamy, SK; Elçi, A. Un modelo de predicción eficaz para la tasa de abandono de cursos en línea. En t. J. Educación a Distancia Tecnología **2020**, 18, 94–110. [Referencia cruzada]
2. Wang, W.; Yu, H.; Miao, C. Modelo profundo para la predicción de la deserción en los MOOC. En Actas de la 2da Conferencia Internacional sobre Criptografía, Seguridad y Privacidad, Beijing, China, 6–9 de julio de 2017; págs. 26–32. [Referencia cruzada]
3. Prenkaj, B.; Velardi, P.; Stilo, G.; Distant, D.; Faralli, S. Una encuesta de enfoques de aprendizaje automático para la predicción de la deserción estudiantil en Cursos en Línea. *Cómputo ACM. sobrev.* **2020**, 53, 1–34. [Referencia cruzada]

4. Queiroga, EM; Lopes, JL; Kappel, K.; Aguiar, M.; Araújo, RM; Muñoz, R.; Villarroel, R.; Cechinel, C. Un enfoque analítico de aprendizaje para identificar estudiantes en riesgo de deserción: un estudio de caso con un curso de educación técnica a distancia. *aplicación ciencia* **2020**, 10, 3998. [\[Referencia cruzada\]](#)
5. Lu, X.; Wang, S.; Huang, J.; Chen, W.; Yan, Z. ¿Qué determina la deserción en los MOOC? En *Actas de la Conferencia Internacional sobre Sistemas de Bases de Datos para Aplicaciones Avanzadas*, Suzhou, China, 27–30 de marzo de 2017; págs. 316–327. [\[Referencia cruzada\]](#)
6. Yang, B.; Qu, Z. Extracción de funciones y análisis de efectos de aprendizaje para usuarios de MOOCs basados en minería de datos. *Educ. ciencia Teoría Practica*. **2018**, 18, 1138–1149. [\[Referencia cruzada\]](#)
7. Moreno-Marcos, PM; Alario-Hoyos, C.; Muñoz-Merino, PJ; Kloos, Predicción de CD en MOOC: una revisión e investigación futura *Direcciones. Trans. IEEE. Aprender. Tecnología* **2019**, 12, 384–401. [\[Referencia cruzada\]](#)
8. Mubarak, AA; Cao, H.; Zhang, W. Predicción de la deserción temprana de los estudiantes basada en sus registros de interacción en el entorno de aprendizaje en línea. *Interactuar. Aprender. Reinar*. **2020**. [\[Referencia cruzada\]](#)
9. Jin, C. Modelo de predicción de abandono de estudiantes de MOOC basado en características de comportamiento de aprendizaje y optimización de parámetros. *Interactuar. Aprender. Alrededor de*. **2020**. [\[Referencia cruzada\]](#)
10. Driik, M.; Munk, M.; Skalka, J. Identificación de cambios en el comportamiento de las partes interesadas de EVA a lo largo del tiempo mediante la minería de patrones frecuentes. *Acceso IEEE* **2021**, 9, 23795–23813. [\[Referencia cruzada\]](#)
11. Shaun, R.; Panadero, JD; Inventado, PS *Minería de datos educativos y análisis de aprendizaje*; Springer: Berlín/Heidelberg, Alemania, 2014; Capítulo 4; págs. 61–75.
12. Siemens, G.; Baker, RSJD *Análisis de aprendizaje y minería de datos educativos*. En *Actas de la 2da Conferencia Internacional on Cryptography, Security and Privacy*, Vancouver, BC, Canadá, 29 de abril a 2 de mayo de 2012; págs. 252–254.
13. Alamri, A.; Al Shehri, M.; Cristea, A.; Pereira, DF; Oliveira, E.; Shi, L.; Stewart, C. Predicción de la deserción de los MOOC utilizando solo dos características fácilmente obtenibles de las actividades de la primera semana. En *Actas de la Conferencia Internacional sobre Sistemas Inteligentes de Tutoría*, Kingston, Jamaica, 3–7 de junio de 2019; Springer Science and Business Media LLC: Berlín/Heidelberg, Alemania, 2019; Volumen 11528, págs. 163–173.
14. Romero, C.; Ventura, S.; Panadero, R.; Pechenizkiy, M. *Manual de minería de datos educativos* (Chapman & Hall/CRC Data Mining and Serie de Descubrimiento del Conocimiento); CRC Press: Boca Raton, FL, EE. UU., 2010; ISBN 1584888784.
15. Skalka, J.; Driik, M. Evaluación automatizada y unidades de microaprendizaje como predictores de estudiantes en riesgo y resultados de estudiantes en los Cursos de Introducción a la Programación. *aplicación ciencia* **2020**, 10, 4566. [\[Referencia cruzada\]](#)
16. Ifenthaler, D.; Yau, JY-K. Utilizando análisis de aprendizaje para apoyar el éxito del estudio en la educación superior: una revisión sistemática. *Educ. Tecnología Res. desarrollo* **2020**, 68, 1961–1990. [\[Referencia cruzada\]](#)
17. Driik, M.; Munk, M. Comprensión de las tendencias basadas en el tiempo en la elección del tipo de actividad de aprendizaje de las partes interesadas mediante el uso predictivo Modelos. *Acceso IEEE* **2018**, 7, 3106–3121. [\[Referencia cruzada\]](#)
18. Hämmäläinen, W.; Vinni, M. Classiers para la minería de datos educativos. En *Manual de Minería de Datos Educativos*; Romero, C., Ventura, S., Pechenizkiy, M., Baker, RSJD, Eds.; CRC Press: Boca Raton, FL, EE. UU., 2010; páginas. 57–74.
19. Lang, C.; Siemens, G.; Sabio, A.; Gasevic, D. (Eds.) *Manual de análisis de aprendizaje*; SOLAR: Beaumont, AB, Canadá, 2017.
20. Kloft, M.; Stiehler, F.; Zheng, Z.; Pinkwart, N. Predicción del abandono de MOOC durante semanas utilizando métodos de aprendizaje automático. En *Actas del taller EMNLP 2014 sobre el análisis de la interacción social a gran escala en los MOOC*, Doha, Qatar, 25 al 29 de octubre de 2014; págs. 60–65. [\[Referencia cruzada\]](#)
21. Uden, L. Tecnología de aprendizaje para la educación en la nube. común computar información *ciencia* **2015**, 533, 43–53. [\[Referencia cruzada\]](#)
22. Baneres, D.; Rodríguez-González, ME; Serra, M. Un sistema de predicción de retroalimentación temprana para estudiantes en riesgo dentro de un curso de educación superior de primer año. *Trans. IEEE. Aprender. Tecnología* **2019**, 12, 249–263. [\[Referencia cruzada\]](#)
23. Kennedy, G.; Ataüd, C.; De Barba, P.; Corrin, L. Predicción del éxito. En *Actas de la Quinta Conferencia Internacional sobre Interacción Tangible, Incrustada y Corporizada*, Poughkeepsie, NY, EE. UU., 16 al 20 de marzo de 2015; págs. 136–140. [\[Referencia cruzada\]](#)
24. Olivé, DM; Huynh, DQ; Reynolds, M.; Dougiamas, M.; Wiese, D. Un marco de aprendizaje supervisado: uso de la evaluación para identificar a los estudiantes en riesgo de abandonar un MOOC. *J. Cómputo. Alto. Educ.* **2019**, 32, 9–26. [\[Referencia cruzada\]](#)
25. Benko, L.; Reichel, J.; Munk, M. Análisis del comportamiento de los estudiantes en un entorno de aprendizaje virtual en función de las evaluaciones de los estudiantes. En *Actas de la 13.ª Conferencia internacional sobre tecnologías y aplicaciones de aprendizaje electrónico emergentes (ICETA) de 2015*, Stary Smokovec, Eslovaquia, 26 y 27 de noviembre de 2015; IEEE: Piscataway, Nueva Jersey, EE. UU.; págs. 1 a 6.
26. Heródoto, C.; Rienties, B.; Borooa, A.; Zdrahal, Z.; Hlosta, M.; Naydenova, G. Implementando analíticas de aprendizaje predictivo a gran escala. En *Actas de la Séptima Conferencia Internacional sobre Análisis y Conocimiento del Aprendizaje*, Vancouver, BC, Canadá, 13–17 de marzo de 2017; págs. 267–271. [\[Referencia cruzada\]](#)
27. Márquez-Vera, C.; Cano, A.; Romero, C.; Noamán, AYM; Fardoun, HM; Ventura, S. Predicción temprana de la deserción usando datos minería: un estudio de caso con estudiantes de secundaria. *Sistema experto* **2016**, 33, 107–124. [\[Referencia cruzada\]](#)
28. Charitopoulos, A.; Rangoussi, M.; Koulouriotis, D. Sobre el uso de métodos de computación blanda en la investigación de análisis de aprendizaje y minería de datos educativos : una revisión de los años 2010–2018. En *t. J. Artif. Intel. Educ.* **2020**, 30, 371–430. [\[Referencia cruzada\]](#)
29. Romero, C.; Espejo, PG; Zafra, A.; Romero, JR; Ventura, S. Minería de uso web para predecir calificaciones finales de estudiantes que usan Cursos Moodle. *computar aplicación Ing. Educ.* **2010**, 21, 135–146. [\[Referencia cruzada\]](#)
30. Rastrollo-Guerrero, J.L.; Gómez-Pulido, J.A.; Durán-Domínguez, A. Analyzing and Predicting Students' Performance by Means del aprendizaje automático: una revisión. *aplicación ciencia* **2020**, 10, 1042. [\[Referencia cruzada\]](#)

31. Xing, W.; Chen, X.; Stein, J.; Marcinkowski, M. Erratum: Corrección de "Predicción temporal de abandonos en MOOC: Alcanzar la fruta al alcance de la mano a través de la generalización del apilamiento" (Computers in Human Behavior (2016) 58 (119–129)(S074756321530279X)(10.1016/j.chb.2015.12).007)). *computar Comportamiento humano* **2017**, 66, 409. [\[Referencia cruzada\]](#)
32. Youssef, M.; Mohamed, S.; Hamada, EK; Wafaa, BF Un enfoque predictivo basado en la selección eficiente de características y la competencia de algoritmos de aprendizaje: caso de abandono de los estudiantes en los MOOC. *Educ. información Tecnología* **2019**, 24, 3591–3618. [\[Referencia cruzada\]](#)
33. Obonya, J.; Kapusta, J. Identificación de actividades importantes para la enseñanza de lenguajes de programación mediante árboles de decisión. En *Actas de la 12.ª Conferencia Científica Internacional sobre Aprendizaje a Distancia en Informática Aplicada (DIVAI)*, Štúrovo, Eslovaquia, 2–4 de mayo de 2018; Turcani, M., Balogh, Z., Munk, M., Kapusta, J., Benko, L., Eds.; Kluiwert: Sturovo, Eslovaquia, 2018; págs. 481–490.
34. Hagedoorn, TR; Spanakis, G. Perfiles temporales de cursos masivos abiertos en línea para la predicción de la deserción. En *Actas de la 29.ª Conferencia internacional sobre herramientas con inteligencia artificial (ICTAI) del IEEE de 2017*, Boston, MA, EE. UU., 6 al 8 de noviembre de 2017; págs. 231–238.
35. Lacave, C.; Molina, AI; Cruz-Lemus, JA Learning Analytics para identificar los factores de abandono de los estudios de Ciencias de la Computación a través de Redes bayesianas. *Comportamiento información Tecnología* **2018**, 37, 993–1007. [\[Referencia cruzada\]](#)
36. Doleck, T.; Lemay, DJ; Basnet, RB; Bazelaïs, P. Análisis predictivo en educación: una comparación de marcos de aprendizaje profundo. *Educ. información Tecnología* **2020**, 25, 1951–1963. [\[Referencia cruzada\]](#)
37. Ali, M. PyCaret: una biblioteca de aprendizaje automático de código abierto y código bajo en Python. 2021. Disponible en línea: <https://pycaret.readthedocs.io/en/latest/index.html> (consultado el 10 de febrero de 2021).
38. Dietterich, TG Pruebas estadísticas aproximadas para comparar algoritmos de aprendizaje de clasificación supervisada. *Cómputo neuronal*. **1998**, 10, 1895–1923. [\[Referencia cruzada\]](#) [\[PubMed\]](#)
39. Ferguson, R.; Cooper, A.; Drachsler, H.; Kismihok, G.; Boyer, A.; Tammets, K.; Monés, AM Analítica de aprendizaje. En *Actas de la Quinta Conferencia Internacional sobre Interacción Tangible, Incrustada y Corporizada*, Poughkeepsie, NY, EE. UU., 16 al 20 de marzo de 2015; págs. 69–72.