Técnicas de *machine learning* aplicadas a la evaluación del rendimiento y a la predicción de la deserción de estudiantes universitarios, una revisión.

Edmanuel Cruz[1], Marvin González[1], José Carlos Rangel[1]
[1]RobotSIS, Universidad Tecnológica de Panamá, Panamá.
edmanuel.cruz, marvin.gonzalez2, jose.rangel {@utp.ac.pa}
DOI: 10.33412/pri.v13.1.3039

Resumen: *En los últimos años, técnicas de Inteligencia Artificial (IA) como el aprendizaje automático o* Machine Learning *(ML) y el Aprendizaje profundo o* Deep Learning *(DL) han impactado de forma positiva el avance de distintos campos del conocimiento; entre ellos, la educación. La educación es un importante motor de todas las sociedades, permite a los individuos ser más productivos y resolver problemas con mayor efectividad, aplicando generalmente enfoques creativos. En la educación se ha utilizado las técnicas de ML para distintas tareas, entre ellas, la predicción de deserción y ayuda al rendimiento del estudiante. En este estudio analizaremos los trabajos más relevantes en estos campos, otorgando una perspectiva de cómo han influenciado los algoritmos de ML y DL en la educación. La búsqueda de los artículos ha sido realizada utilizando la herramienta de búsqueda proporcionada por* Google Scholar. *Las búsquedas se hicieron usando las palabras claves:* Student Dropout, Student Performance Prediction y Machine Learning. *Los artículos fueron seleccionados por* relevance *(relevancia).*

Palabras claves: Inteligencia Artificial, Aprendizaje Automático, Aprendizaje Profundo, mejoramiento estudiantil, deserción estudiantil, Predicción del rendimiento de los estudiantes.

Title: Machine Learning Techniques Applied to Evaluate the Performance and Dropout Prediction of University Student's, A Review.

Abstract: In recent years, Artificial Intelligence (AI) techniques such as Machine Learning (ML) and Deep Learning (DL) have positively impacted the advancement of various fields of knowledge, including education. Education is an important engine of all societies; education allows individuals to be more productive and solve problems more effectively by generally applying creative approaches. In education, the above-mentioned AI techniques have been used for different tasks, among them, student dropout prediction and help to the student's performance. In this study we will analyze the most relevant works in these fields, giving a perspective of how ML and DL algorithms have influenced education.

Key words: Artificial Intelligence, Machine Learning, Deep Learning, Student Dropout, Student Improvement, Student Performance Prediction.

Tipo de artículo: revisión.
Fecha de recepción: 8 de junio de 2021
Fecha de aceptación: 31 de enero de 2022.

1. Introducción

Tanto para los países como para las personas existe un vínculo directo entre el acceso a una educación de calidad y el desarrollo social y económico. Todos los países, independientemente de su nivel de riqueza, se beneficiarían de una mejor y mayor cobertura en educación. Según la Organización para la Cooperación y el Desarrollo Económicos (OECD, por sus siglas en inglés), si se proporcionara a todos los niños acceso a la educación y a las aptitudes necesarias para participar plenamente en la sociedad, el Producto Interno Bruto aumentaría en un promedio del 28% anual en los países de ingresos bajos y del 16% anual en los países de ingresos altos durante los próximos 80 años [1]. Por otro lado, tenemos los avances en Inteligencia Artificial (IA), la cual es una herramienta de gran alcance que permite a la gente repensar la forma en que integramos la información, analizamos los datos y utilizamos los conocimientos resultantes para mejorar la toma de decisiones. Hoy se encuentra como un ente que transforma todos los ámbitos de la vida [2].

La inteligencia artificial cuenta con distintas ramas, que abarcan distintos campos de estudios. Los autores de [3], realizaron un resumen esquemático de las principales ramas de la inteligencia artificial (IA), incluidos los métodos de aprendizaje automático (ML). En la Figura 1 se puede apreciar dicho resumen esquemático. En este estudio nos concentraremos en las técnicas de *Machine Learning* y *Deep Learning.*

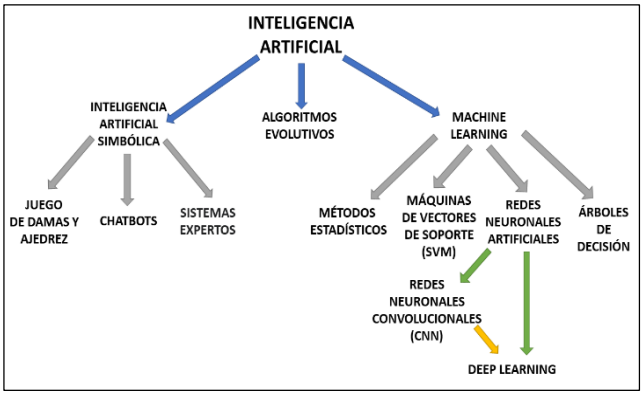

Figura 1. Resumen esquemático de las principales ramas de la inteligencia artificial (IA).

En la actualidad, técnicas de IA como el *Machine Learning* están transformando la educación y cambiando fundamentalmente la enseñanza, el aprendizaje y la investigación. Los educadores están utilizando el ML para detectar de forma temprana a los estudiantes con dificultades y tomar medidas para mejorar el éxito y la retención[1] del estudiante en el sistema.

En el caso de la deserción, esta ocurre por distintos factores. Primeramente, puede ser voluntaria, en cuyo caso, aunque se deserte de una carrera puede transferirse a otra carrera o a otra universidad. Segundo, tenemos la deserción involuntaria, la cual conlleva casi siempre a la deserción total. Esta se da por factores socioeconómicos, motivaciones e intereses del estudiante. Aunque la transferencia depende de las motivaciones del estudiante, también puede ser influenciada por las condiciones dentro de la universidad [4]. Con respecto a la predicción del rendimiento de los estudiantes, se puede decir que es uno de los temas más importantes para los contextos de aprendizaje como las escuelas y las universidades, ya que ayuda a diseñar mecanismos eficaces que mejoran los resultados académicos y evitan el abandono escolar, entre otras cosas [5].

En este sentido, muchos investigadores han enfocado sus esfuerzos en aplicar distintas técnicas de IA en líneas de investigación como la predicción de deserción y la ayuda al rendimiento del estudiante.

Aunque existen distintas revisiones de la literatura, tanto en predicción de deserción [6][7][8][9][10] como en el rendimiento y mejoramiento del desempeño estudiantil [11][12], son pocos los trabajos que revisan ambos temas de investigación.

Así, pues, motivados por las actuales tendencias, en el presente documento se examinan y resumen las prometedoras y desafiantes investigaciones sobre la predicción de deserción y la ayuda al rendimiento del estudiante, principalmente universitario, usando técnicas de ML.

La búsqueda de los artículos ha sido realizada utilizando la herramienta de búsqueda proporcionada por *Google Scholar*. Las búsquedas se hicieron usando las palabras claves: *Student Dropout, Student Performance Prediction y Machine Learning*. Los artículos fueron seleccionados por su relevancia y su año de publicación, empleando los filtros proporcionados por *Google Scholar*. Dichos filtros fueron: artículos publicados desde 2016 hasta el 2021, ordenar búsqueda en orden de relevancia y cualquier idioma. Finalmente, los artículos con mayor relevancia a la temática fueron utilizados en este estudio.

El resto del artículo se estructura de la siguiente manera: En la sección 2 proporcionaremos la definición del problema de deserción y presentaremos los enfoques recientes que hacen uso de ML o DL para afrontar este problema. En la sección 3 veremos el tema de la predicción del desempeño del estudiante y la mejora del rendimiento de este. Finalmente, en la sección 4 se expondrán las conclusiones.

2. Predicción de deserción estudiantil

El fenómeno de la deserción universitaria se da en todas las universidades del mundo y sus efectos se sienten tanto a nivel económico, como a nivel de la sociedad y personal de los estudiantes que abandonan la universidad. Los factores que conducen al abandono de la universidad pueden ser de carácter social (antecedentes de los estudiantes, nivel de ingresos, etc.), psicopedagógico (inadecuado nivel académico de formación, inconsistencia entre la formación previa y los estudios universitarios, falta de servicios de asesoría, etc.) y personal (poca adaptabilidad a la universidad, bajos niveles de inteligencia socioemocional, etc.) [13].

El problema de la predicción de deserción estudiantil suele ser abordado como un problema de clasificación binaria, donde *0* indica un estudiante que permanece en el curso y *1* representa el estudiante que abandona. Según [14] el nivel de deserción en cursos online es entre 60% a 80% y en cursos presenciales es de alrededor de 40% [15]. Dado el creciente uso de las plataformas virtuales como medio de educación, situación creada por la pandemia ocurrida en el año 2020, analizaremos mayormente la deserción en plataformas virtuales. Según la literatura existen tres tipos de estrategias de predicción: Examen analítico, métodos clásicos de Aprendizaje automático o ML y Aprendizaje profundo o DL [6].

Resumiremos la primera y profundizaremos en los métodos de ML y DL. Cabe destacar que, casi siempre, los estudios de este tipo no toman en cuenta la situación financiera de los estudiantes, debido a que generalmente es información privada. A pesar de esto, estamos conscientes de que la situación financiera influye en la permanencia o no de los estudiantes en las universidades tal y como se destaca en [16].

2.1 Examen analítico

Generalmente en este tipo de exámenes se recogen datos de diversas fuentes y luego se realiza un análisis de correlación entre las características extraídas y la etiqueta de abandono.

Según [6] estos métodos son insensibles al tiempo y no garantizan hallazgos estables, ya que los patrones de abandono pueden cambiar con el tiempo. En este mismo trabajo los autores analizan dos estudios. En el primero [17], los autores recolectaron datos a través de encuestas sobre la deserción escolar que se produce en la educación superior en el campo de las ciencias de la computación, durante dos años. Luego examinaron la correlación entre el abandono de los estudiantes y los perfiles de los estudiantes. Proporcionan información estadísticamente significativa sobre la correlación de la presentación de asignaciones y la educación previa con la decisión de abandonar la escuela. Finalmente concluyen que los estudiantes que abandonan los estudios estiman erróneamente la carga de estudio mientras trabajan. Igualmente, una minoría de ellos siente que su tutor no les ayudó a comprender el material del curso y a completar sus tareas.

En un segundo estudio [18], los autores desarrollaron y probaron modelos predictivos usando los datos históricos de la actividad combinada con otras fuentes en los entornos de aprendizaje virtuales para tres módulos de la Open University (OU). Esto reveló que es posible predecir el fallo de los estudiantes observando el comportamiento de los cambios en las actividades en los entornos virtuales cuando es comparado con su propio comportamiento en ocasiones anteriores o cuando es comparado con estudiantes categorizados con comportamiento similar. Los autores también investigaron sobre el método de datos GUHA (*General Unary Hypothesis Automaton*) que es un método de generación automática de hipótesis basado en datos empíricos. Este método genera hipótesis a partir de los datos basados en los parámetros iniciales los cuales son: confianza, la cual delinea la probabilidad de que una hipótesis generada clasifique correctamente las etiquetas; el soporte, que es el porcentaje mínimo de las reglas que se ajusta a la regla generada, y el número máximo de antecedentes, que corresponde al número de literales que se producen en la parte izquierda de la implicación. Los autores muestran que no presentar la cuarta evaluación lleva a un abandono completo del curso. Sin embargo, los resultados son exactos cuando se aplican a través de diferentes presentaciones del módulo [6].

2.2 Métodos Clásicos de Machine Learning

El *Machine Learning* es una rama evolutiva de los algoritmos computacionales que están diseñados para emular la inteligencia humana aprendiendo del entorno [19]. Los modelos generados por Machine Learning han demostrado un gran éxito en el aprendizaje de patrones complejos que les permiten hacer predicciones sobre datos no observados [20]. Las técnicas basadas en ML se han aplicado con éxito en diversos campos que van desde el reconocimiento de patrones, la visión por ordenador, las finanzas, la ingeniería de naves espaciales, el entretenimiento, la biología computacional, aplicaciones médicas y por supuesto la educación. A continuación, describiremos en orden cronológico, desde 2016 hasta 2021, los trabajos que utilizan técnicas clásicas de ML para predicción de la deserción estudiantil.

2.2.1 Trabajos presentados desde el año 2016 hasta el 2021

En [21], se utilizó un dataset extraído del registro de la Universidad de Washington (UW). Los datos contienen la información demográfica (raza, sexo, fecha de nacimiento, condición de residente e identificación como hispano), información sobre el ingreso a la universidad (resultados de los exámenes SAT y ACT, si están disponibles) y registros completos de las transcripciones (clases tomadas, tiempo en que se tomaron, calificaciones recibidas y áreas de especialización declaradas) de todos los estudiantes del sistema de la Universidad de Washington (UW) (que consiste en un campus principal en Seattle y dos campus satélites: Bothell y Tacoma). Tomaron muestras al azar de la mayoría para crear un conjunto equilibrado de datos compuesto por 32,538 estudiantes. En el mapeo de características, la etnia, el género y la condición de residente eran variables categóricas en las que cada estudiante sólo pertenecía a una única categoría y la inclusión en las categorías era mutuamente excluyente entre las variables. Cada posible etnia (6 en total), género (3 en total), y estatus de residente (7 en total) fueron mapeados a través de variables ficticias. Para realizar sus experimentos utilizaron tres modelos de ML: *regularized logistic regression* (regresión logística regularizada), *k-nearest neighbors* (k- vecinos más cercanos) y *Random Forests* (bosques aleatorios) para predecir la variable binaria de abandono sobre las características descritas anteriormente. En todos los experimentos, analizan el rendimiento con un 30% de los datos, seleccionados al azar. Con el 70% restante de los datos, se utilizó una validación cruzada con k=10 para ajustar los parámetros del modelo (por ejemplo, la fuerza de regularización para la regresión logística, el número de vecinos en kNN, y la profundidad del árbol en bosques aleatorios). Según los autores la regresión logística regularizada proporcionó las mejores predicciones. Predecir el número de trimestres que tardan en terminar los cursos antes de retirarse dio resultados marginales, ya que las predicciones tenían un error cuadrático medio (RMSE) de unos 5 trimestres de matriculación.

Los autores en [22] proponen utilizar el algoritmo ID3 [23][24] de J. Ross Quinlan. Este algoritmo, el ID3, utiliza un enfoque voraz de arriba hacia abajo para construir un árbol de decisiones. Explicado de forma más simple, el enfoque descendente significa que empezamos a construir el árbol desde la cima y el enfoque voraz significa que en cada iteración seleccionamos la mejor característica del momento para crear un nodo. En este trabajo los autores han mejorado el tradicional algoritmo ID3 mediante el uso de la entropía de Rényi [25]. Esta combinación se utiliza como un nuevo criterio para construir el árbol de decisión y predecir el abandono de los estudiantes universitarios. Para este estudio empírico, que consta de 32 variables, se utilizó un conjunto de datos de 240 muestras recogidas al azar mediante una encuesta en una universidad situada en la India. Los autores reportaron que su algoritmo obtuvo un 97% de exactitud sobre los datos utilizados versus 92% que obtuvo un modelo de árbol de decisión tradicional.

En este trabajo [26], los autores propusieron un sistema de predicción de la deserción en plataformas MOOC (cursos masivos abiertos en línea), mediante un algoritmo no supervisado que utiliza datos históricos con el objetivo de predecir antes de que ocurra. Para llevar a cabo su trabajo utilizaron los datos del curso de Estructuras de Datos y Algoritmos de la Universidad de Pekín en Coursera. El curso dura 14 semanas y consiste en vídeos de conferencias, pruebas, tareas de programación y un foro de discusión. El sistema propuesto hace uso de *Random Forest* para clasificar los datos y se utilizó F-1 Score como medida de precisión. Basándose en los resultado de los experimentos del sistema de predicción de abandono escolar, los autores hacen varias sugerencias para ayudar a mejorar la gestión del curso desde la perspectiva de la prevención del abandono, tales como ofrecer a los estudiantes más oportunidades de realizar pruebas y tareas, prolongar el periodo de realización de las tareas calificadas, animar a los estudiantes a participar en los foros de discusión y diseñar pruebas en vídeo para dividir cada vídeo en fragmentos cortos.

Los autores, en [27], han investigado el nivel práctico de precisión que puede lograrse con un predictor automático de abandono del MOOC. Para desarrollar su enfoque, hacen uso de una arquitectura de clasificación de los detectores de regresión logística con regularización de L2, lo que equivale a una red neuronal de dos capas. Los experimentos y análisis de este estudio se basan en datos de 40 MOOCs de HarvardX. Para medir la precisión de los clasificadores de abandono, los autores utilizaron la métrica de la Curva de las características de funcionamiento del receptor (AUC). La propia curva de características operativas del receptor (ROC) traza la tasa positiva verdadera frente a la tasa positiva falsa del clasificador entrenado. Un punto destacable de este estudio es la implementación de una red neuronal prealimentada (*feed-forward network*) que les permite extraer características demográficas básicas y de flujo de clics más profundas con los cuales obtuvieron mejores resultados. Los resultados de su investigación sugieren que la precisión de los clasificadores que se entrenan con datos que se recogen sólo después de que un curso haya terminado y que, por lo tanto, no son utilizables en el propio MOOC suelen ser varios puntos porcentuales más altos que los clasificadores que se entrenan en otros MOOC. Según los autores, esto subraya la importancia de una cuidadosa estimación de la precisión antes de llevar a cabo una intervención a gran escala.

En la propuesta [28], los autores introducen una metodología para predecir la deserción estudiantil utilizando el algoritmo de clasificación Naïve Bayes en el lenguaje R. Este estudio también examina la razón de la deserción de los estudiantes en un estado temprano y predecir si el estudiante abandonará o no. Por lo tanto, la recopilación de datos juega un papel importante en este trabajo. Los datos recogidos son evaluados por las diversas técnicas de preprocesamiento de datos. Los datos contienen información sobre 54 atributos diferentes de cada estudiante.

Los datos recopilados por diversos recursos muestran que muchos factores como los académicos, los demográficos, los psicológicos, los de salud, etc. juegan un papel importante en el abandono escolar. En esta investigación utilizan ciertas metodologías como la técnica de cálculo, la identificación de factores y la preparación de un cuestionario de encuesta [29].

La arquitectura es descrita de la siguiente manera:

- El estudiante es el usuario que interactúa con el sistema a través de la API de Google. La API consiste en el conjunto de encuesta, pre-encuesta y post-encuesta.
- La pre-encuesta se utiliza para encuestar al estudiante del primer semestre y la encuesta posterior es utilizada para encuestar al estudiante del tercer semestre.
- Las dos se utilizan para distinguir el punto de vista de un estudiante que se une a la universidad (pre-encuesta) y el estudiante que continúa en la universidad (post-encuesta).
- Los formularios de Google son usados para hacer el informe de la encuesta y luego son almacenados en la base de datos.
- Los datos almacenados son procesados mediante la técnica de preprocesamiento de datos. La reducción de la dimensionalidad [30] tiene un papel importante en la arquitectura reduciendo los atributos de la colección de conjuntos de atributos
- Luego los datos de los atributos se almacenan en la base de datos y los datos se evalúan usando *InfoGainAttributeEval*.
- El paso final es la predicción. Muestra si el correspondiente estudiante cometerá abandono de estudios en un formato de sí o no.

Según los resultados presentados, este modelo ayudará a identificar al estudiante que va a abandonar el curso registrado.

En [31], los autores describen los resultados de un caso de estudio de análisis de datos educativos enfocado en la detección de deserción de estudiantes de pregrado de Ingeniería de Sistemas (SE) luego de 7 años de matrícula en una universidad colombiana. El dataset utilizado en este trabajo proviene de 802 estudiantes matriculados en el Programa de Ciencias de la Computación en una universidad privada en Bogotá, Colombia.

En este estudio se comparan los resultados de Árboles de Decisión (*Decision Trees*), Regresión Logística (*Logistic Regression*) y Naïve Bayes para proponer la mejor opción. Sus resultados experimentales mostraron que los algoritmos simples logran niveles confiables de precisión para identificar el abandono estudiantil. Además, evalúan el servicio de Watson Analytics para establecer la usabilidad del servicio para un usuario no experto. Watson Analytics es un servicio inteligente para analizar y visualizar datos para descubrir rápidamente patrones y significado en los datos, sin tener ningún conocimiento previo. Watson Analytics utiliza el descubrimiento de datos guiado, el análisis predictivo automatizado y las capacidades cognitivas para interactuar con los datos y obtener hallazgos que comprenda. Los resultados experimentales mostraron que el mejor AUC se logró mediante el modelo de árbol de decisión (0,94), por lo que esta precisión podría ser lo suficientemente segura como para ayudar a la detección temprana de abandono. En los resultados muestran que el rendimiento de los cursos de ingeniería de sistemas está correlacionado con el rendimiento de los cursos de física y matemáticas.

Los autores de [32] proponen aplicar algoritmos de ML para predecir la deserción escolar. Para ello hacen uso de los datos proporcionados por la organización Uwezo[2], la cual recoge datos sobre el nivel de educación del este de África. Este dataset contiene información escolar entre 2009 a 2015 de Kenia, Tanzania y Uganda. En este estudio los autores toman 4 algoritmos de aprendizaje supervisado los cuales son evaluados con 3 métricas: media geométrica, F-score, media geométrica ajustada. Utilizan como caso de estudio los datos de Tanzania. Estos datos fueron procesados y luego comparados entre ellos, además se evalúa el efecto de utilizar un modelo base y otro ajustando los hiperparámetro. La regresión logística y perceptrón multicapa son los que presentan mejores resultados cuando se utiliza la técnica de over-sampling.

En [15], los autores realizaron sus pruebas con dos algoritmos: Perceptrón multicapa y RBF (*radial basis function*)

Networks. Los datos utilizados fueron recolectados a través de encuestas a través de Formularios de Google. Para ellos fueron encuestados 2060 estudiantes de primero a cuarto año matriculados y no graduados de carreras de administración y humanidades de la Universidad pública de Ecuador entre los años 2014 a 2017. En el perceptrón multicapa se usó 60% de los datos para entrenamiento (1602 casos), 30% para validación (801) y 10% (267) como test. Y en el caso de la red RBF se usó 70% para entrenamiento y 30% para test. Con el modelo de Perceptrón Multicapa obtuvieron una exactitud de 96,3% en entrenamiento y 98,6% en Test. Mientras que con la Red RBF obtuvieron 96,8% en entrenamiento y 98,1% en test. Siendo el Perceptrón Multicapa el mejor modelo en cuanto a generalización de los datos se refiere.

En [33] utilizan regresión logística y arboles de decisión para predecir la deserción de estudiantes del Instituto de Tecnología de Karlsruhe. En este estudio se muestran cuáles son las principales causas de deserción en las universidades de Alemania. Para el desarrollo de las pruebas se utilizó datos de estudiantes de Ingeniería Industrial del Instituto de Tecnología de Karlsruhe que empezaron a estudiar entre el 2007 y 2012 periodo de otoño. En la selección de características, se tomaron 487 muestras de exámenes, notas, fechas, resultados y número de intentos. Los autores reflejan en sus resultados que es posible entrenar modelos de ML sobre datos puramente académicos sin necesitad de evaluar otros factores que afecten la privacidad de los datos. Entre las desventajas de este modelo entrenado tenemos que no puede ser transferible a otra universidad o institución ya que se entrenó sobre datos específicos de una sola facultad (Ing. Industrial); sin embargo, la técnica que se utiliza si puede ser replicada. A partir de estos datos puede obtenerse una exactitud (*accuracy*) de hasta el 95%, aunque este porcentaje baja a medida que se analizan datos de semestres más avanzados. Esto se debe principalmente al desbalance que hay en los datos.

En el enfoque presentado en [34], los autores utilizan *Random Forest* y Arboles de decisión como algoritmos de ML para predecir la deserción estudiantil en los estudiantes de Ingeniería Informática. Utilizan datos de 206 estudiantes de primer año del programa de Ing. Informática de la Universidad de Santiago de Chile entre los años 2012 - 2016, excepto del 2015. De estos datos, 146 pertenecen a la clase "*retention*" que representa a los estudiantes que pasaron al segundo año de estudio y 60 que pertenecen a la clase "*dropout*" que significa que desertaron. Incluye 40 *datapoints* (características) que van desde datos académicos, socioeconómicos y demográficos. Como test se usan datos del 2017 - 2018. De las 40 características se hace una selección de 7 utilizando *Random Forest*, donde 6 de ellas corresponden a factores académicos y una a un factor socioeconómico. Después que se hace la selección de características, se utiliza un árbol de decisión y se usan todos los datos del 2012 2016 como conjunto de entrenamiento. Los resultados los presentan en matrices de confusión con un 97,2% de exactitud.

Los autores en [35] presentan un modelo para explicar y predecir el abandono universitario, y diseñar acciones para reducirlo. Para ello utiliza un algoritmo k-means para clasificar y definir los patrones de rendimiento, y las predicciones para los nuevos estudiantes se realizan mediante un modelo de máquina de vectores de soporte (del inglés *Support Vector Machines*, SVM). Los datos utilizados en el estudio fueron recolectados de los estudiantes que se inscribieron en dos períodos de admisión de la Universidad Tecnológica Indoamérica de Ambato, Ecuador. Según los autores, los resultados permiten a las instituciones y al profesorado centrarse en los grupos de alto riesgo durante los primeros trimestres y modificar su comportamiento de aprendizaje futuro.

2.2.2 Otros Estudios

El estudio [36] está enfocado en la deserción estudiantil en preparatoria, sin embargo, presenta un enfoque interesante. Los autores proponen una metodología y un algoritmo de clasificación específico para descubrir modelos de predicción comprensibles de la deserción escolar lo antes posible. El dataset utilizado en este trabajo proviene de 419 estudiantes matriculados en la Unidad Académica Preparatoria de la Universidad Autónoma de Zacatecas (UAPUAZ) en México. En este estudio detallan 3 experimentos llevados a cabo para predecir la deserción en diferentes etapas del curso, para seleccionar los mejores indicadores de deserción y para comparar su algoritmo propuesto con algunos algoritmos clásicos de clasificación conocidos. El algoritmo propuesto por los autores fue llamado ICRM2 puesto que es una versión mejorada de su previo algoritmo llamado Interpretable *Classification Rule Mining* (ICRM) [37]. El algoritmo se compone de tres fases, en la primera fase, se crea un conjunto de reglas que exploran los dominios de atributos. En la segunda fase, el algoritmo itera para encontrar las reglas de clasificación y construye el clasificador. Finalmente, la tercera fase optimiza la precisión del clasificador. La metodología del ICRM fue adaptada para centrarse en la predicción de los estudiantes que abandonan los estudios antes de tiempo, donde hay menos información disponible sobre los estudiantes. Además, el procedimiento de generación de reglas del algoritmo ICRM2 fue adaptado para generar conjuntos de reglas centradas en la clase de datos desequilibrada (estudiantes que abandonan la escuela). ICRM2 genera dos conjuntos de reglas: el primero muestra las reglas que predicen el éxito del estudiante, mientras que el segundo predice el abandono del estudiante. Este algoritmo fue comparado con algoritmos tradicionales de ML o variantes de estos utilizando el software Weka[3]. Los algoritmos usados fueron: *Bayesian classifier*, Naïve Bayes, SVM [38], *Sequential minimal optimization* (SMO) [39], *K-nearest neighbours classifier*, *Classification rules [40], Decision trees*. Los autores reportaron que su algoritmo fue capaz de predecir la deserción de los estudiantes en las primeras 4 − 6 semanas del curso.

2.3 Métodos de *Deep Learning*

Actualmente los cursos masivos abiertos en línea (MOOC) se han vuelto muy populares y más con la pandemia, sin embargo, la tasa de abandono es también alta. Cómo predecir eficazmente el estado de abandono de los estudiantes en los MOOC para intervenir lo antes posible se ha convertido en un tema de actualidad. En este sentido, en [41] han propuesto un algoritmo que utiliza el word2vector para codificar, luego utiliza una red neuronal convolucional (CNN) para extraer características, luego utiliza una red de memoria de larga a corto plazo (LSTM) para combinar las características temporales de cada vector de entrada, y finalmente utiliza bosques aleatorios (RF) para predecir. El algoritmo puede extraer automáticamente las series temporales y aprovechar al máximo las ventajas de los algoritmos anteriores para mejorar el rendimiento del modelo. Los experimentos con conjuntos de datos comunes muestran que el AUC del modelo mejora significativamente en comparación con los algoritmos existentes. Este estudio [14], el problema de predicción de deserción se considera como predicción de series de tiempo. Se evalúan dos MOOCs ofrecidos por Coursera y edX con una RNN y una LSTM, para ello utilizan MATLAB. Este tipo de cursos la tasa de deserción está entre el 60-80%, por lo cual los datos se encuentran desbalanceados, pero esta vez con mayoría a la deserción. Para entrenar el modelo secuencial, se consideró como periodo de tiempo las semanas. En una semana un estudiante realiza distintas actividades dentro del curso como lecturas, exámenes, videos, etc. Esto se utiliza como entrada X a la red para predecir una salida Y en un tiempo t. La salida Y representa si deserta o no. Los autores reportan buenos resultados tratando el tema como un problema de clasificación de secuencias y aplicando modelos temporales para resolverlo.

En [42] los autores tienen como objetivo predecir si un estudiante de un curso en línea desertara en los siguientes 10 días. Según este trabajo, un problema común en sistemas de predicción es la parte de ingeniería de características que se suele hacer de forma manual y para cada plataforma o curso suele crearse un dataset no generalizable. Se utiliza una arquitectura a la cual llaman ConRec Network, este consiste en 2 redes trabajando conjuntamente. La primera parte es una CNN para extraer las características de los datos en crudo, a partir de estos para cada intervalo de tiempo forma matrices de características que son el input de una RNN que combina la información para predecir si el estudiante va a desertar o no. El dataset utilizado es el KDD Cup 2015[43]. Este dataset está en formato de texto y contiene distintos atributos como el tipo de actividad que realiza un estudiante, su ID y lo trabajan como si fuera un problema de Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés) para luego hacer *one-hot vector*. Cabe mencionar que los autores compararon el modelo de su red con algoritmos de ML como SVM, *Random Forest*, arboles de decisión y entre otros. Para la implementación de su red utilizaron herramientas de software como Python y la librería Theano. Como métricas de evaluación utilizaron precisión, recall, F1-score y AUC. Obtuvieron resultados alrededor del 88%. Su principal ventaja está en que no extraen manualmente las características relevantes, esto le permite adaptarlos a otras plataformas o cursos.

El estudio presentado en [44] hace a referencia a la alta deserción existente en cursos virtuales basándose en estudios como los presentados por [45]. Aquí proponen utilizar una red FWTS-CNN. Extraen las características más importantes de un registro de las actividades de los estudiantes a través de árboles de decisión, se les da un peso a estas de acuerdo con su importancia, se construyen matrices por intervalos de tiempo que luego se utilizan como entrada para una red CNN. Los datos utilizados son los del dataset KDD Cup 2015 que provienen de la plataforma XuetangX. Se usa el 80% para entrenamiento y 20% para test. El resultado lo comparan con modelos bases de regresión lineal, SVM, *Random Forest* y otros, usando las métricas de Accuracy, Precision, Recall y F1-Score. Se obtienen resultados por encima del 86% en la predicción de las tasas de deserción para nuevos cursos.

El enfoque presentado en [46] buscaba crear una plataforma para predecir la deserción de los estudiantes mediante datos socioeconómicos. Este sistema debía ser capaz de ser accedido desde distintos dispositivos. El dataset utilizado está compuesto por datos socioeconómicos de cuatro cursos de ingeniería (civil, computación, mecatrónica y telecomunicaciones) del Instituto Federal de Educación, Ciencia y Tecnología de Ceará (IFCE). Campus Fortaleza, Brasil.

Los datos se obtuvieron de los formularios de registro de los estudiantes entre los años 2008-2019, teniendo en total 1549 registros de los cuales 1318 culminaron los cursos y 231 desertaron.

Se obtuvieron los datos de los registros y se extrajeron 7 características: Género, Edad, Etnia, Índice de desarrollo humano por vecindario, Ingreso familiar, Escuela de procedencia, Distancia desde la universidad. Se reprocesaron las características ya que algunas son de tipo categóricas y luego se normalizaron entre -1 y 1. Para realizar este preprocesamiento se utilizó Python para la parte de preprocesamiento y predicción en una aplicación web. Para la parte de servicio web se utilizó Java y la información se almacenó en PostgreSQL. HTML5, CSS3 para *front-end* de la aplicación web, y JSON para el intercambio de datos. Para el entrenamiento de los algoritmos se dividieron los datos en 80% para entrenamiento y 20% para test. Cabe mencionar que los datos se dividieron de forma aleatoria. Se utilizaron algoritmos de ML como Árbol de Decisión, Regresión logística, SVM con kernel RBF, KNN, MLP y DNN. Las métricas utilizadas para la evaluación del modelo fueron: accuracy, F1 score, recall, y precisión. Se obtuvo 99,34% accuracy, 99,34% F1 score, 100% recall, y 98,69% de precisión usando arboles de decisión. Además, la DNN utilizada obtuvo un resultado por encima al 90%. Este resultado ha demostrado que la utilización de información relativa al entorno socioeconómico de los estudiantes permite realizar predicciones con alto valor de certeza en cuanto a la posibilidad de deserción estudiantil. De igual manera, la amplia variedad de algoritmos permite intuir que este tipo de datos es adaptable a diversos métodos de predicción.

3. Evaluación del rendimiento de los estudiantes

El principal objetivo de cualquier institución educativa es ofrecer la mejor experiencia y conocimiento educativo a los estudiantes. En este sentido, identificar a los estudiantes que necesitan apoyo adicional y tomar las acciones apropiadas para mejorar su desempeño juega un papel importante en el logro de esa meta.

En esta sección veremos las distintas técnicas de ML utilizadas para predecir el desempeño de los estudiantes.

3.1 Predicción del Desempeño del Estudiante

Dentro de los objetivos que se tiene de la inclusión de métodos de análisis de información de los estudiantes de educación superior, está el hecho de predecir el desempeño que tendrá un estudiante en un curso tomando en cuenta su desenvolvimiento actual. Esta rama surge con el objetivo de realizar una detección temprana y evitar la deserción o el fracaso.

El uso de ML ha permitido explorar diferentes métodos en diferentes tipos de datos, obteniendo resultados prometedores y estableciendo la base para estos experimentos. En los principios de la década los algoritmos utilizados se enfocaban solo en ML tradicional los cuales no incluían procesamiento con DL, ni acceso a mayores volúmenes de información como se observa en la revisión publicada en [47], donde se presenta un análisis de los trabajos enfocados en los algoritmos utilizados, predominan en este caso las redes neuronales, SVM y algoritmos de regresión.

Dicha revisión implementa una evaluación de la correlación entre el desempeño de los estudiantes y el estado psicológico de estos. En los últimos años son diversos los métodos aplicados para este análisis los cuales se pueden ver agrupados en la siguiente tabla 1.

Tabla 1. Métodos aplicados en la predicción del rendimiento estudiantil.

Algoritmo	Aplicado en:
Naïve Bayes	[48] [49]
SVM	[50] [48] [49]
Redes Neuronales	[51] [50] [48] [49]
Regresión Logística	[51] [48] [49]
Arboles de Decisión	[51] [50] [48] [49]
KNN	[48] [49]
Bagging y Boosting	[48] [49]

Siguiendo el desarrollo tanto del ML como de los estudios de desempeño, se inician estudios donde se incluyen otras áreas de la Inteligencia Artificial como el Procesamiento del Lenguaje Natural (NLP). Este enfoque se presenta en [52], en el cual los autores utilizan texto no estructurado proveniente de preguntas aplicadas a los estudiantes de MOOCs. La información o datos fue además filtrada mediante una serie de filtros demográficos enfocados en la depuración de los datos, algunos de estos filtros son la fluidez en el idioma inglés, el inicio en el curso en las 2 primeras semanas y la intención del participante en terminar el curso.

El estimador creado se desarrolló utilizado el enfoque de regresión logística con penalización, la cual se determina empíricamente y minimiza el error de predicción durante el proceso de entrenamiento utilizado enfoques de validación cruzada (*cross-validation*). Las pruebas demuestran que NLP permite la predicción del desempeño de los estudiantes en el curso, pero se reconocen limitaciones del método como lo son que el enfoque puede ser aplicado a pequeños conjuntos de estudiantes y también que son necesarios grandes cantidades de datos para lograr un análisis más efectivo.

Los experimentos presentados en [51] evalúan el impacto que puede producir la psicología del estudiante en la predicción de sus notas. Para la recopilación de datos se aplican cuestionarios basados en el Inventario de estrategia de estudio y aprendizaje [53] (LASSI, por sus siglas en inglés) de tal manera que se pueda conocer el nivel de motivación del estudiante y la manera en la cual utilizan los recursos de aprendizaje. Para el dataset se realizó una encuesta de 98 preguntas a un grupo de 150 estudiantes, además se incluía información sobre calificaciones preuniversitarias, calificación de ingreso a la universidad, el promedio de 6 semestres anteriores y el tipo de personalidad. El modelo propuesto emplea regresión a través de una red neuronal para predecir la calificación del estudiante y árboles de decisión para predecir si el estudiante aprobará o no la asignatura. Los resultados del estudio logran establecer que la inclusión de aspectos psicológicos favorece la estimación del desempeño y permite al cuerpo docente aplicar los correctivos a fin de evitar futuros fracasos de los estudiantes.

Dentro de los enfoques encontrados un elemento cambiante es el tipo de datos que se utiliza para la creación de los modelos en el caso del enfoque presentado en [50], se utilizan clasificadores como arboles de decisión, redes neuronales y SVM. En este caso, se buscaba determinar si características relacionadas a la conexión a internet podrían influir en el desempeño académico de los estudiantes. La información fue recabada mediante las cuentas de acceso a internet dentro de la universidad de cada estudiante. Luego del análisis y filtrado de los datos se obtuvieron más de 20 millones de registros para 4,000 muestras de estudiantes. Dichas muestras se analizaron con un coeficiente de correlación de Spearman, para posteriormente entrenar los modelos de predicción. Como resultado del estudio se determinó que el desempeño se ve asociado positivamente con la frecuencia de conexión a internet, mientras que estaba negativamente asociado al volumen del tráfico durante las conexiones.

Existen otros trabajos cuyos datos se enfocan en el análisis concreto de los resultados de los estudiantes para un grupo de materias. En este caso, las pruebas presentadas en [48] [49] se enfocan recolectar y procesar información relacionada a las asignaturas que forman parte del concepto STEM (*Science, Technology, Engineering and Mathematics*). Estos se enfocan en la evaluación de varios modelos y estrategias de predicción para clasificar a un estudiante dentro de 3 posibles grupos tomado en cuenta los resultados obtenidos por estos en exámenes cortos, tareas, asignaciones y exámenes finales, obteniendo un total de 538 estudiantes. Sus categorías clasifican el desempeño como

Bueno, Aceptable y En Riesgo. El estudio agrupa los elementos para cada estudiante según la semana en la cual fueron asignadas, esto se realiza de forma incremental, es decir, la primera agrupación contiene datos de la semana 1 a la semana 3, el segundo grupo contendrá información de la semana 1 a la semana 6, el grupo 3 de la semana 1 a la semana 9 y el grupo 4 de la semana 1 a la semana 12. Utilizando estos grupos se realizan 4 predicciones que permiten estimar la situación del estudiante en diversos momentos, lo cual ayuda al instructor a detectar cualquier posible problema existente en los diferentes momentos del curso. En este estudio se implementan tanto clasificadores como ensambles de estos para determinar el mejor modelo posible. Se crean, por lo tanto, 7 modelos predictivos empleando Regresión Logística, KNN, *Random Forest, Redes Neuronales, Gradient Boosting y Adaptative Boosting.* Los resultados arrojan que la mejor ejecución se alcanzó utilizando *Random Forest,* el cual se genera utilizando el esquema de *Bagging*.

La predicción del desempeño de los estudiantes ha encontrado diversas maneras de evaluar el problema presentando mayores diferencias en el tipo de información que se analiza para realizar la predicción. Entre estas se puede mencionar los comentarios, aspectos psicológicos, información del estatus del estudiante en el curso y el desempeño en ciertas temáticas específicas. La información incluye captación de datos mediante diversos métodos y escalas que permiten una variabilidad en las respuestas y por lo tanto hacen el contenido del dataset más diverso. En esta línea de investigación se pueden mencionar como métodos más utilizados las redes neuronales y los árboles de decisión, pero a pesar de ser los más evaluados, no son los algoritmos que obtienen los mejores resultados en los experimentos. Siendo que cada trabajo utiliza un tipo de dato diferente, los mejores resultados para cada uno son obtenidos por diferentes algoritmos entre estos SVM y *Random Forest* que han sido utilizados y comparados en los diferentes experimentos.

3.2 Análisis de la retroalimentación en cursos universitarios

En el apartado anterior contemplamos la utilización de técnicas de ML enfocados en mejorar el rendimiento de los estudiantes mediante la predicción de una calificación o la probabilidad de que un estudiante apruebe un curso. Como segunda línea predominante que busca la mejora del rendimiento se encuentran los trabajos que se enfocan en el análisis de comentarios (*feedback*) dejados por los estudiantes referentes a un curso, a docentes o metodologías empleadas. Los estudios en su mayoría integran NLP a los cuales se unen también algoritmos de ML o minería de datos con la intención de predecir, principalmente, la polaridad que presenta dicha retroalimentación.

La información para esta línea de investigación es recolectada mayormente mediante las encuestas de evaluación de asignaturas y docentes que se aplican al final de los periodos académicos. Esta información sirve también como punto de partida para investigaciones relacionadas con el enfoque de Evaluación de la Enseñanza por los Estudiantes (SET, por sus siglas en inglés). Un resumen de los diferentes métodos usados en los trabajos expuestos se encuentra en el tabla 2.

Tabla 2. Métodos aplicados en el análisis de comentarios

Algoritmo	Aplicado en:
Naïve Bayes	[54] [55] [56] [57]
SVM	[54] [58] [55] [57]
Árboles de Decisión	[58] [55] [57]
Redes Neuronales	[54]
KNN	[54]
K-Means	[55]
Entropía Máxima	[56]
Redes Neuronales Profundas	[56]

Dentro de los aspectos que se extraen de los comentarios de retroalimentación se puede mencionar la polaridad del comentario lo cual ayuda a crear un análisis amplio de los comentarios recibidos. Trabajos como [54] buscan la polaridad mediante la utilización de un dataset de comentarios que contiene la observación personal de cada alumno relativa a los exámenes, el proceso de enseñanza, el contenido de los módulos y los recursos de laboratorio. El dataset contiene información para 6 cursos y un total de aproximadamente 13, 000 comentarios. Las palabras son vectorizadas utilizando TF-IDF (*Term Frequency-Inverse Document Frequency*) y luego son empleados para entrenar los clasificadores. Los autores utilizan la aplicación RapidMiner (para preprocesado y entrenamiento), con el cual entrenan clasificadores aplicando KNN, SVM. Naïve Bayes y una Red Neuronal. Los resultados señalan que para esta representación vectorizada de los comentarios un clasificador de Naïves Bayes obtiene los mejores resultados.

Un enfoque similar al anterior se puede encontrar en los experimentos de [58], donde se emplea de igual manera TF-IDF para la representación vectorial del texto tokenizado, con la intención de clasificar los comentarios en 3 posibles estados: positivo, negativo o neutral. En este caso el preprocesamiento de los datos se realizó mediante la librería NLTK de Python. El dataset está conformado por 1203 comentarios extraídos de un portal universitario y clasificado manualmente. En la tabla 3 se presentan ejemplos de estos comentarios y su polaridad asignada. El preprocesamiento del dataset involucra de igual manera un diccionario que contiene la polaridad de cada palabra. Para este estudio se empleanpalabras clasificadas y relacionadas al ámbito académico; este enfoque recibe el nombre de *Lexicon based Method*, por lo cual, la propuesta se centra en un enfoque híbrido entre un diccionario y métodos de ML. Utilizando este diccionario se puede asignar la polaridad a un comentario tomando en cuenta la polaridad predominante entre las palabras que le componen. Luego del preprocesamiento y vectorización de las características se entrenan clasificadores utilizando SVM y *Random Forest*, siendo este ultimó el clasificador que obtuvo mejores resultados. El estudio también compara su enfoque con herramientas de análisis de sentimientos disponibles en la web como *Text Analytics API, Alchemy*

Language API y Aylien Text API, donde el modelo híbrido supera los resultados obtenidos con estas herramientas.

Tabla 3. Comentarios de ejemplo en el dataset (en el idioma original).

Comentario	Etiqueta
timings are very odd such courses should not be offered at such late timings, as for programming PERSON NEED FRESH MIND. Till the time of our class we r all dead tired and sleepy.	Negativo
She is a very hard-working instructor, actually helps us a lot however the course is way too irrelevant for ACF students	Positivo
Give more programming assignments and enhance the level of the course to include critical thinking and solutions as it is required in CS research	Neutral

El procesamiento de comentarios de retroalimentación se ha enfocado de diversas maneras entre estos el etiquetado POS (*Part-Of-Speech*) el cual se centra en la clasificación y etiquetado de palabras de acuerdo con su parte en la oración. Dentro de estos estudios podemos mencionar [55] quienes aplican POS a un conjunto de comentarios recolectados desde un API de Twitter.

Luego del primer análisis, se asigna cada opinión a un aspecto relativo al entorno educativo, por ejemplo, algunos de los aspectos fueron: enseñanza, instalaciones y transporte. Para ello se aplica la medición Omiotis que establece la relación semántica entre los aspectos y la oración, esto con la intención de asignar cada oración a un aspecto. El estudio asigna un sentimiento a cada oración mediante el módulo de análisis de sentimientos de R y K-means ejecutado en Weka y como elemento final determina la polaridad de la oración mediante Naive Bayes, el cual superó en este aspecto a los Árboles de Decisión y SVM.

Con el auge del DL, el campo del análisis de comentarios ha obtenido nuevas maneras de evaluar la información y por lo tanto se realizan comparativas entre métodos tradicionales y métodos y los nuevos enfoques. En este caso, el estudio [56] compara un clasificador entrenado utilizando algoritmos tradicionales como Nave Bayes y Máxima Entropía, para compararlos con algoritmos de Deep Learning, en específico las Redes Neuronales Recurrentes (LSTM) y Redes Neuronales Recurrentes Bidireccionales (BLSTM). Usan en este estudio el UIT-VSFC: *Vietnamese Students' Feedback Corpus for Sentiment Analysis* [59] que contiene más de 16, 000 comentarios recolectados en durante 3 años. Estos comentarios están clasificados en positivos, negativos y neutrales. La información fue codificada utilizando POS y DEP (*Dependency on Relation*) para los algoritmos tradicionales y WORD2VEC para los algoritmos de RNN. Sus resultados determinan que las redes BLSTM obtienen una mejor precisión en comparación con los algoritmos tradicionales.

Como se ha apreciado en los últimos estudios presentados los enfoques son muy parecidos, pero la mayor diferencia se puede encontrar en la manera en la cual se preprocesan y se extraen las características de los comentarios en los datasets. En este los experimentos presentados en [57] emplean la función CountVectorizer disponible en la librería SciKitLearn de Python.

Los comentarios son recolectados mediante un formulario de Google, estos comentarios son clasificados inicialmente mediante la herramienta VADERSentiment disponible para Python, la cual asignará una polaridad a cada respuesta (positiva, negativa, neutral). Seguidamente se extraerán las características de las oraciones y junto con su polaridad se utilizarán para entrenar diversos clasificadores. Las características se calcularán utilizando Count-Vectorizer, con la intención de representar la información en un formato adecuado para ser utilizado por los algoritmos. En este estudio se entrenaron modelos utilizando los algoritmos SVM, *Random Forest* y Naive Bayes Multinomial. Siendo el clasificador Naive Bayes Multinomial quien obtuvo mejores resultados en los diferentes experimentos.

Dentro de este bloque de estudios encontramos el estudio de [57] cuyo objetivo se centra en analizar los sentimientos de los estudiantes y emociones de los estudiantes mediante técnicas de minería de sentimientos. El estudio busca no solo enfocarse en respuestas individuales de cada estudiante, sino en crear una herramienta que permita un análisis a gran escala de una masiva cantidad de información proveniente de las universidades. Esta propuesta utiliza la Asignación Latente de Dirichlet (LDA, por sus siglas en inglés), para realizar el análisis bajo un enfoque no supervisado, del conjunto de datos utilizado en la experimentación. Mediante LDA se identificarán los tópicos presentes en los comentarios y estos serán enlazados con los tópicos identificados en otros estudios con la intención de que esta nueva información pueda ser utilizada por educadores y administrativos, ya que se utilizaría un vocabulario común. Para los estudios se utilizaron datos recolectados en dos periodos académicos previos, los cuales fueron enviados por correo electrónico y contenían preguntas como las siguientes:

- Otro comentario referente al curso (por ejemplo, maneras de mejorar el aprendizaje durante el curso)
- ¿Qué factores influyeron en cuánto invertí en mi aprendizaje?
- ¿Qué factores influyeron en mi nivel de motivación?

Al final del proceso de recolección se obtuvieron 6,087 respuestas utilizadas en el estudio, estas estaban escritas en finlandés. Como resultado del análisis del conjunto de datos, se identificaron 6 tópicos los cuales serían asignados a los diferentes comentarios emitidos por los estudiantes. Dichos tópicos fueron relacionados con construcciones similares en la literatura existente, mediante un análisis cualitativo. De igual manera se presenta una evaluación cuantitativa donde se analizan los tópicos con la escala de Likert, utilizando métodos estadísticos. Estos resultados apoyan el enfoque propuesto y validan un método nuevo para validar los tópicos que se puedan identificar en un conjunto de datos de esta naturaleza.

De igual manera las pruebas presentadas en [60] utilizan un dataset de 30,000 muestras basadas en comentarios y sentimientos de estudiantes relativos a la apreciación de un curso. En este estudio se enfocan en el desarrollo de un modelo de clasificación utilizando algoritmos de ML tradicional, como lo son

SVM, Naïve Bayes y una Red Neuronal. Sus resultados determinan que, para el tipo de dato contenido en el conjunto de entrenamiento, la red neuronal obtiene un mejor desempeño.

La temática relacionada al análisis de los comentarios de los estudiantes esta mayormente orientada a extraer la polaridad de dichos comentarios, en los estudios consultados se puede observar un proceso común a muchos otros aplicados en otras áreas, siendo la principal diferencia entre cada enfoque, la manera o métodos que se emplean para representar los datos y el preprocesamiento de estos (POS, WORD2VEC, TF-IDF, Diccionarios Léxicos). No se puede observar en el área un dataset común o popular, sino que cada trabajo se enfoca más en recolectar su propia fuente de datos para su problema y objetivo específico. De entre los algoritmos empleados, se puede mencionar Naïve Bayes como el enfoque que mayormente se ha aplicado en la solución del problema estudiado, el cual obtiene valores de F1-Score entre 80-90%, representando los mejores resultados en la mayoría de los trabajos en que ha sido aplicado.

Cabe resaltar que los enfoques con redes neuronales reportan obtener resultados satisfactorios; sin embargo, son hasta el momento utilizados por menos autores en esta línea de investigación. Se debe tomar en cuenta que, si bien existe un objetivo para esta área, es común que se apliquen diversos algoritmos de IA o ML a pasos previos que para la preparación o análisis de los datos que se usarán en el proceso de entrenamiento principal. Por lo cual se puede decir que los diferentes algoritmos para el aprendizaje automático encuentran cabida en los diferentes niveles de procesamiento. De igual manera, la revisión nos indica que el NLP y sus múltiples herramientas es el área de la Inteligencia Artificial que ha permitido el avance en los procesos de análisis de retroalimentación para centros de estudios.

Conclusión

En los últimos años la creciente tasa de abandono de los estudiantes en un curso registrado ha sido una gran amenaza para muchas instituciones educativas o universidades. El estudiante entra en la institución con muchas expectativas y sueños, sin embargo, sus expectativas no se cumplen o ciertos factores como la demografía afectan y hacen que abandonen sus estudios. Como hemos visto, una gran variedad de técnicas de Inteligencia Artificial ha sido adaptadas o desarrolladas para predecir este tipo de situaciones, pero para que estas técnicas sean realmente aprovechadas, deben ser exitosamente adoptadas por las instituciones educativas. Por otro lado, tenemos la aplicación de algoritmos de Inteligencia Artificial dirigidos a tratar de ayudar por retroalimentación a los estudiantes en el mejoramiento de su rendimiento. En este documento se presenta una visión general de soluciones en ambos temas. Estas soluciones se benefician de los nuevos avances en Inteligencia Artificial ya que estos aumentan la efectividad de los sistemas dirigidos a las temáticas antes mencionadas, permitiéndoles mejores resultados. A pesar del activo desarrollo de la tecnología de predicción de deserción estudiantil y ayuda al mejoramiento del rendimiento de los estudiantes, todavía hay trabajo por hacer.

De hecho, las soluciones actuales no proporcionan soluciones ideales a todas las necesidades de los estudiantes e instituciones, pero los resultados son muy prometedores.

Agradecimientos

Edmanuel Cruz y José Carlos Rangel son apoyados con fondos del Sistema Nacional de Investigación de la SENACYT.

Referencias

[1] OECD, E, A. Hanushek, and L. Woessmann. Universal Basic Skills. 2015
[2] D. M. West and J. R. Allen. How artificial intelligence is transforming the world, Apr 2020.
[3] F. Galbusera, G. Casaroli and T. Bassani Artificial intelligence and machine
learning in spine research. JOR SPINE, 2(1), p.e1044. 2019
[4] B. M Kehm, M. Rode Larsen, and H. Bjørnøy Sommersel. Student dropout from universities in europe: A review of empirical literature. Hungarian Educational Research Journal, 9(2):147–164, 2019
[5] J. L Rastrollo-Guerrero, J. A Gómez-Pulido, and A. Durán - Domínguez. Analyzing and predicting students' performance by means of machine learning: A review. Applied Sciences, 10(3):1042, 2020.
[6] B. Prenkaj, P. Velardi, G. Stilo, D. Distante, and S. Faralli. A survey of machine
learning approaches for student dropout prediction in online courses. ACM Computing Surveys (CSUR), 53(3):1–34, 2020.
[7] F. Dalipi, A. Shariq Imran, and Z. Kastrati. Mooc dropout prediction using machine learning techniques: Review and research challenges. In 2018 IEEE Global Engineering Education Conference (EDUCON), pages 1007–1014. IEEE, 2018.
[8] S Thulasi Bharathi. Analysis on massive open online course (mooc) dropout
prediction using machine learning techniques-the state of the art.
[9] N. Mduma, K. Kalegele, and D. Machuve. A survey of machine learning approaches and techniques for student dropout prediction. 2019.
[10] M. Alban and D. Mauricio. Predicting university dropout through data mining:
A systematic literature. Indian Journal of Science and Technology, 12(4):1–12, 2019.
[11] S. Ulfa, R. Bringula, C. Kurniawan, and M. Fadhli. Student feedback on online learning by using sentiment analysis: A literature review. In 2020 6th International Conference on Education and Technology (ICET), pages 53–58, 2020.
[12] M. Edalati. The potential of machine learning and nlp for handling students 'feedback (a short survey), 2020.
[13] C. St˘aiculescu and R. Nastase E. Ramona. University dropout. Causes and
solution. Mental Health: Global Challenges Journal, 1(1):71–75, 2018.
[14] M. Fei and D. Y. Yeung. Temporal models for predicting student dropout in massive open online courses. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pages 256–263. IEEE, 2015.
[15] M. Alban and D. Mauricio. Neural networks to predict dropout at the universities. International Journal of Machine Learning and Computing, 9(2):149–153, 2019.
[16] A. F Cabrera, A. Nora, and M. B Castaneda. The role of finances in the persistence process: A structural model. Research in higher education, 33(5):571–593, 1992.
[17] M. Xenos, C. Pierrakeas, and P. Pintelas. A survey on student dropout rates and dropout causes concerning the students during informatics of the hellenic open university. Computers & Education, 39(4):361–377, 2002.
[18] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek. Improving retention: predic- ting at-risk students by analyzing clicking behavior in a virtual learning environment. In Proceedings of the third international conference on learning analytics and knowledge, pages 145–149, 2013
[19] I. El Naqa and M. J Murphy. What is machine learning? In Machine Learning in Radiation Oncology, pages 3–11. Springer, 2015.
[20] W. James Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu.

Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44):22071–22080, Oct 2019.
[21] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West. Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364, 2016.
[22] S. Sivakumar, S. Venkataraman, and R. Selvaraj. Predictive modeling of student dropout indicators in educational data mining using improved decision tree. Indian Journal of Science and Technology, 9(4):1–5, 2016.
[23] J. R. Quinlan. Discovering rules by induction from large collections of examples. Expert systems in the microelectronics age, 1979.
[24] J. Ross Quinlan. Induction of decision trees. Machine learning, 1(1):81–106, 1986.
[25] A. Renyi et al. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. The Regents of the University of California, 1961.
[26] Y. Chen and M. Zhang. Mooc student dropout: Pattern and prevention. In Proceedings of the ACM Turing 50th Celebration Conference-China, pages 1–6, 2017.
[27] J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley. Delving deeper into mooc student dropout prediction. arXiv preprint arXiv:1702.06404, 2017.
[28] V. Hegde and PP Prageeth. Higher education student dropout prediction and analysis through educational data mining. In 2018 2nd International Conference on Inventive Systems and Control (ICISC), pages 694–699. IEEE, 2018.
[29] M. Xenos, C. Pierrakeas, and P. Pintelas. A survey on student dropout rates and dropout causes concerning the students during informatics of the hellenic open university. Computers & Education, 39(4):361–377, 2002.
[30] V. Hegde. Dimensionality reduction technique for developing undergraduate student dropout model using principal component analysis through r package. In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pages 1–6. IEEE, 2016.
[31] B. Perez, C. Castellanos, and D. Correal. Applying data mining techniques to predict student dropout: a case study. In 2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI), pages 1–6. IEEE, 2018.
[32] N. Mduma, K. Kalegele, and D. Machuve. Machine learning approach for reducing student's dropout rates. 2019.
[33] L. Kemper, G. Vorhoff, and B. U. Wigger. Predicting student dropout: A machine learning approach. European Journal of Higher Education, 10(1):28–47, 2020.
[34] F. A Bello, J. K´ohler, K. Hinrechsen, V. Araya, L. Hidalgo, and J. L. Jara. Using machine learning methods to identify significant variables for the prediction of first-year informatics engineering students' dropout. In 2020 39th International Conference of the Chilean Computer Science Society (SCCC), pages 1–5. IEEE, 2020.
[35] A. F. Núñez - Naranjo, M. Ayala-Chauvin, and G. Riba-Sanmart´ı. Prediction of university dropout using machine learning. In Álvaro Rocha, Carlos Ferras, Paulo Carlos López- López, and Teresa Guarda, editors, Information Technology and Systems, pages 396–406, Cham, 2021. Springer International Publishing.
[36] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura. Early dropout prediction using data mining: a case study with high school students. Expert Systems, 33(1):107–124, 2016.
[37] A. Cano, A. Zafra, and S. Ventura. An interpretable classification rule mining algorithm. Information Sciences, 240:1–20, 2013.
[38] C. Cortes and V. Vapnik. Support-vector networks. Machine Learning, 20(3):273–297, September 1995.
[39] J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.
[40] W. W Cohen. Fast effective rule induction. In Machine learning proceedings 1995, pages 115–123. Elsevier, 1995.
[41] L. Cai and G. Zhang. Prediction of moocs dropout based on wclsrt model. In 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), volume 5, pages 780–784, 2021.
[42] W. Wang, H. Yu, and C. Miao. Deep model for dropout prediction in moocs. In Proceedings of the 2nd International Conference on Crowd Science and Engineering, pages 26–32, 2017.
[43] KDD Cup. Kdd cup 2015: Predicting dropouts in mooc, 2015.
[44] Y. Zheng, Z. Gao, Y. Wang, and Q. Fu. Mooc dropout prediction using fwts-cnn model based on fused feature weighting and time series. IEEE Access, 8:225324–225335, 2020.
[45] T. R Liyanagunawardena, P. Parslow, and S. Williams. Dropout: Mooc participants' perspective. 2014.
[46] F. A da S Freitas, F. FX Vasconcelos, S. A Peixoto, M. Mehedi Hassan, M Dewan, V. Hugo C de Albuquerque, et al. Iot system for school dropout prediction using machine learning techniques based on socioeconomic data. Electronics, 9(10):1613, 2020.
[47] R. R. Halde. Application of machine learning algorithms for betterment in education system. In 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), pages 1110–1114, 2016.
[48] M. Aly and M. Rashedul Hasan. Improving stem performance by leveraging machine learning models. In 2019 International Conference in Frontiers in Education: CS and CE, 2019.
[49] M. Hasan and M. Aly. Get more from less: A hybrid machine learning framework for improving early predictions in stem education. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI), pages 826–831, 2019.
[50] X. Xu, J. Wang, H. Peng, and R. Wu. Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Computers in Human Behavior, 98:166 – 173, 2019.
[51] R. R. Halde, A. Deshpande, and A. Mahajan. Psychology assisted prediction of academic performance using machine learning. In 2016 IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT), pages 431–435, 2016.
[52] C. Robinson, M. Yeomans, J. Reich, C. Hulleman, and H. Gehlbach. Forecasting student achievement in moocs with natural language processing. In Proceedings of the Sixth International Conference on Learning Analytics and Knowledge, LAK 16, page 383–387, New York, NY, USA, 2016. Association for Computing Machinery.
[53] C.E. Weinstein and D.R. Palmer. LASSI-HS User's Manual. H & H Pub., 1990.
[54] V. Dhanalakshmi, D. Bino, and A. M. Saravanan. Opinion mining from student feedback data using supervised learning algorithms. In 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC), pages 1–5, 2016.
[55] M. Sivakumar and U. S. Reddy. Aspect based sentiment analysis of students opinion using machine learning techniques. In 2017 International Conference on Inventive Computing and Informatics (ICICI), pages 726–731, 2017.
[56] P. X. V. Nguyen, T. T. T. Hong, K. V. Nguyen, and N. L. Nguyen. Deep learning versus traditional classifiers on Vietnamese students' feedback corpus. In 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), pages 75–80, 2018.
[57] D. Deeksha Dsouza, D. P Nayak Deepika, E. Jenisha Machado, and ND Adesh. Sentimental analysis of student feedback using machine learning techniques. IJRTE, ISSN, pages 2277– 3878.
[58] Z. Nasim, Q. Rajput, and S. Haider. Sentiment analysis of student feedback using machine learning and lexicon-based approaches. In 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), pages 1–6, 2017.
[59] K. V. Nguyen, V. D. Nguyen, P. X. V. Nguyen, T. T. H. Truong, and N. L. Nguyen. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In 2018 10th International Conference on Knowledge and Systems Engineering (KSE), pages 19–24, 2018
[60] S. Katragadda, V. Ravi, P. Kumar, and G. J. Lakshmi. Performance analysis on student feedback using machine learning algorithms. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pages 1161–1163, 2020.
[61] https://aws.amazon.com/es/education/ml-in-education/?nc1=h_ls
[62] https://www.twaweza.org/go/uwezo-datasets
[63] https://www.cs.waikato.ac.nz/ml/weka/