DOI: https://doi.org/10.56124/encriptar.v5i10.0050

MODELOS PREDICTIVOS APLICADOS EN LA EDUCACIÓN: CASOS ABANDONO DE ESTUDIO

PREDICTIVE MODELS APPLIED IN EDUCATION: CASES OF DROPPING OUT OF STUDY

Cedeño-Valarezo Luis

Grupo de Investigación SISCOM, Carrera de Computación, ESPAM MFL. Calceta, Ecuador.
Correo: lcedeno@espam.edu.ec

Morales-Carrillo Jessica

Grupo de Investigación SISCOM, Carrera de Computación, ESPAM MFL. Calceta, Ecuador.
Correo: jmorales@espam.edu.ec

Quijije-Vera Carlos Pierre

Grupo de Investigación SISCOM, Carrera de Computación, ESPAM MFL. Calceta, Ecuador.
Correo: cquijije@espam.edu.ec

Palau-Delgado Sandro Antonio

Grupo de Investigación SISCOM, Carrera de Computación, ESPAM MFL. Calceta, Ecuador.
Correo: spalau@espam.edu.ec

RESUMEN

El propósito de esta investigación es analizar datos de una revisión de artículos científicos basados en modelos predictivos empleados en la educación, con especificidad en casos de abandono de estudio con el objetivo de identificar el modelo más eficiente según la frecuencia de uso. Se empleó la metodología de revisión sistemática aplicando un metaanálisis, partiendo con la definición de palabras clave, luego, se integraron criterios como la especificación de la técnica y el tipo de aprendizaje de un determinado modelo. Finalmente, se realizaron pruebas estadísticas en base a la precisión de cada uno. Se evidenció que los árboles de decisión obtuvieron una precisión media de 86.49% con una desviación estándar de 9% en 53 casos encontrados. Además, los modelos de redes neuronales y random forest alcanzaron valores de precisión media de 89.18% y 91.33%, desviación estándar de 5,90% y 3,08% en 7 y 8 casos respectivamente.

Palabras claves: Deserción estudiantil, Repetición estudiantil, Minería de datos, Modelo predictivo.

Fecha de recepción: 05 de mayo de 2022; **Fecha de aceptación:** 20 de junio de 2022; **Fecha de publicación:** 09 de julio de 2022.

ABSTRACT

The purpose of this research is to analyze data from a review of scientific articles based on predictive models used in education, with specificity in cases of study abandonment to identify the most efficient model according to the frequency of use. The systematic review methodology was used applying a meta-analysis, starting with the definition of keywords, then criteria such as the specification of the technique and the type of learning of a certain model were integrated. Finally, statistical tests were performed based on the precision of each one. It was evidenced that the decision trees obtained a mean precision of 86.49% with a standard deviation of 9% in 53 cases found. In addition, the neural network and random forest models reached mean precision values of 89.18% and 91.33%, standard deviation of 5.90% and 3.08% in 7 and 8 cases respectively.

Keywords: Student desertion, Student repetition, Data mining, Predictive model.

1. INTRODUCCIÓN

La Inteligencia Computacional (IC) es un área dentro del campo de la Inteligencia Artificial (IA) que se centra en el diseño de sistemas informáticos inteligentes que imitan la naturaleza y el razonamiento lingüístico humano para resolver problemas complejos (Pandolfi et al., 2017). En este campo de investigación se encuentran tres líneas de desarrollo de sistemas inteligentes que conforman el centro de la IC: la simulación del razonamiento lingüístico humano mediante la teoría de los conjuntos difusos, las redes neuronales que imitan al sistema nervioso, y los algoritmos de optimización (Gonzembach et al., 2021). La aplicación de técnicas de IC para la implementación de modelos predictivos ha aumentado últimamente debido al crecimiento exponencial de datos en distintos campos de las ciencias y las ventajas que conlleva optimizar su proceso de análisis mediante la minería de datos. Oprea, (2020) declara que, durante las últimas dos décadas, la IC se ha desarrollado rápidamente convirtiéndose en uno de los campos más prolíficos de la IA. Es por esto, que la IC se ha transformado en una de las áreas de investigación más activas, especialmente en los últimos años (Abdulrahman et al., 2020). Otros investigadores como Telikani et al., (2020) indican que la Computación Evolutiva (CE) junto con métodos de Minería de Datos por Reglas de Asociación (MDRA) pueden considerarse como una

alternativa a la IC, debido a que los algoritmos aplicados para la predicción en la CE se codifican en función de la solución de un problema; evolucionando constantemente, para brindar soluciones muy cercanas a la óptima, sin embargo, no existe un análisis exhaustivo de las metodologías MDRA evolutivas.

Ahora bien, la minería de datos consiste en el proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados (Guerra et al., 2020). Empleando una amplia variedad de técnicas, puede utilizar esta información para incrementar sus ingresos, recortar costos, mejorar sus relaciones con clientes y reducir riesgos. En este caso la producción de estudios de investigación en base a la minería de datos desde el ámbito educativo, han obtenido beneficios de acuerdo a los procesos de aprendizaje automático que colaboran con la toma de decisiones en las instituciones de educación superior ya que esto hace referencia a las técnicas, herramientas y la investigación diseñados para extraer automáticamente el significado de grandes repositorios de datos generados por o relacionados con las actividades de aprendizaje en los centros educativos.

Las investigaciones sobre el rendimiento académico de los estudiantes aportan en el análisis y diferenciación de los modelos predictivos existentes. También se pueden identificar variables intervinientes para que puedan reproducirse y contribuir a modelos predictivos nuevos y más precisos. La deserción de estudiantes en la educación superior ha constituido un problema medular para las universidades, ya que aproximadamente de 400.000 estudiantes de escuelas politécnicas públicas y privadas del Ecuador, el 26% abandona sus estudios, por lo tanto, debe invertirse en investigaciones relacionadas para predecir casos de abandono de estudio. A través de esta investigación se pretende aportar con evidencia bibliográfica de modelos predictivos precisos, para de esta manera identificar los más eficientes en este tipo de problemática.

2. MATERIALES Y MÉTODOS

Para obtener el objetivo planteado se empleó la metodología de revisión sistemática (Carrizo y Moller, 2018) que consta de tres etapas: definición de la búsqueda, ejecución de la búsqueda y discusión de los resultados.

2.1 Definición para la búsqueda

En la primera etapa se revisó bibliografía específica de artículos científicos en distintas bibliotecas y repositorios digitales, identificando un total de 120 artículos. Específicamente, se tomaron en cuenta artículos desde el 2016, además, la búsqueda se refinó empleando las siguientes palabras clave: Deserción estudiantil, Repitencia estudiantil, Minería de datos, Modelo Predictivo en repositorios como ScienceDirect, IEEExplore, Scielo, Dialnet, Redalyc, SpringerLink, Google Académico.

2.2. Ejecución de la búsqueda

A continuación, se distribuyeron los campos específicos para el proceso de extracción de información de los artículos investigados. En la Tabla 1 se detalla una referencia sobre la formulación de los campos necesarios para realizar la recopilación de información propuesta en la revisión bibliográfica. Se puede agregar también, que los atributos más relevantes de esta matriz el tipo de modelo aplicado y la precisión como métrica de análisis del tipo de aprendizaje. Es necesario señalar que de todos los artículos identificados en primera instancia solo se consideró para el análisis los que contaban con la métrica de precisión.

Siguiendo con el proceso de estudio, las tareas realizadas posteriormente involucraron cálculos de frecuencia, los cuales fueron: En primer lugar, se determinó el número de veces en que cada tipo de modelo predictivo fue utilizado. Luego se calculó la frecuencia absoluta y relativa con respecto a la aplicación general de los modelos identificados que contaban con la precisión. Finalmente, se integraron los valores de la precisión: mínima, máxima y media. Además, se calculó la desviación estándar con el objetivo de corroborar la certeza de un determinado modelo predictivo y asociar este valor con el de la frecuencia absoluta.

***Tabla 1**. Campos que se consideraron en la recopilación de información.*

Campos	Descripción
Año	Año en el que se aprobó y publicó el artículo científico. Solo se investigaron artículos con al menos de 5 años de antigüedad.
Titulo	Nombre con el que se identifica el artículo.
Autor(es)	Aquellos que participaron en la elaboración de los artículos investigados.
Tipo	Tipo de aprendizaje del modelo, dependiendo de los datos disponibles y la tarea que se abordó, se pudo elegir entre distintos tipos de aprendizaje: Aprendizaje supervisado o no supervisado, semi-supervisado y por refuerzo.
Modelo	Grupo de técnicas que, mediante los campos del aprendizaje automático, la recolección de datos históricos, el Big Data y el reconocimiento de patrones, pretende dar una predicción de resultados futuros.
Precisión	Métrica considerada para identificar el modelo más eficiente con respecto a la frecuencia absoluta.

2.3 Discusión de los resultados

En la última etapa, se realizaron pruebas estadísticas específicas con los artículos científicos que se consideraron relevantes de acuerdo con las palabras claves, y tomando en cuenta como punto crítico la implementación de un modelo predictivo como parte de los resultados. De esta forma, se realizaron dos pruebas respectivamente, que involucran actividades de conteo, medidas de tendencia central y dispersión respectivamente:

Análisis de modelos aplicados inicialmente.- Funciona como un punto de partida del estudio de modelos predictivos con el objetivo de indicar cuál es la tendencia de aplicación de consideración en este tipo de problemas.

Demostrar cuál es modelo más eficiente.- Se realizó un análisis estadístico (valoración) basados en la precisión (P), que es una métrica que determina la proporción de verdaderos positivos (VP) entre verdaderos positivos y falsos positivos (FP) predichos, es decir: $P = VP/(VP + FP)$.

3. RESULTADOS Y DISCUSIÓN

3.1 Análisis con respecto al tipo de aprendizaje del modelo

De los 120 artículos identificados inicialmente, solo se tomaron en cuenta 99, para el análisis con respecto a la precisión, ya que contaban dicha métrica. Específicamente 21 artículos del total no tenían el valor de la precisión, ya sea porque el objetivo del artículo no era determinar la validez del modelo propuesto, o, porque el modelo fue evaluado con una métrica distinta.

Gráfico 1. *Tipo de aprendizaje recopilado en los artículos.*

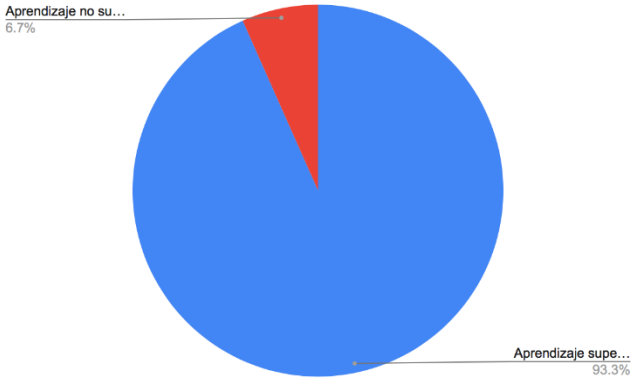
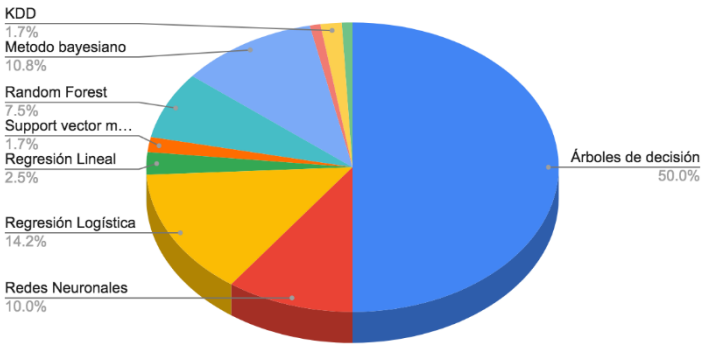


Gráfico 2. *Distribución de los modelos en los artículos.*



De los artículos inicialmente identificados, se empleó mas del 93% con aprendizaje supervisado, y el restante utilizó aprendizaje no supervisado (Gráfico 1).

Los árboles de decisión presentaron una mayor frecuencia de aplicación para este el tipo de problema de deserción estudiantil, ya que probablemente las

variables utilizadas en este modelo se acoplan con mayor eficacia y se puede obtener una mayor precisión con el mismo (Gráfico 2).

3.2 Análisis con respecto a la frecuencia del modelo, identificación de la precisión con su respectiva varianza

Una vez definidos los artículos a revisar, se clasificaron aquellos que cuentan con la métrica de precisión de los que no, ya que este es el requisito principal para ser parte del análisis (tabla 2).

En la tabla 3 se observa que existe un alto índice de precisión en algunos modelos, pero hay que tomar en cuenta la proporción en la que estos modelos fueron implementados, dicha frecuencia es la responsable de que la media de la precisión sea poco confiable.

Tabla 2. *Modelos con métrica precisión frente a los identificados*

Modelos	Identificados	Con métrica precisión
Árboles de decisión	60	53
Redes Neuronales	12	7
Regresión Logística	17	14
Regresión Lineal	3	3
Support vector machine	2	2
Random Forest	9	8
Método bayesiano	13	9
Reglas de asociación	1	1
KDD	2	1
Support Vector Regressor	1	1
Total(T)	**120**	**99**

Se evidencia que a partir de la alta frecuencia con la que se aplican árboles de decisión en esta clase de problemas, la precisión media tendrá un mayor grado de confianza, ya que valores extremos afectan en menor grado la media en comparación con otros modelos en que la frecuencia es baja. En términos más simples, a mayor frecuencia de aplicación del modelo, mayor es la probabilidad de obtener una precisión media más acertada.

En la tabla 3 notamos que los árboles de decisión son los que presentan mayor frecuencia de uso con 54%, con una precisión media del 86.49% y una desviación estándar del 9% a pesar de tener un valor mínimo de 61.00% que está bastante alejado de la media.

Las redes neuronales con una frecuencia de 7% tienen precisión media bastante alta 89.18% y una desviación estándar de 5.90%, sin embargo, surge el inconveniente de estar consideradas únicamente en 7 artículos, por lo que se necesitaría mayor cantidad de casos para poder determinar si la media y su desviación son realmente confiables.

Tabla 3. *Frecuencia y precisión de los modelos analizados.*

Modelos	FA	FR	$\overline{X}$	max	min	σ
Árboles de decisión	53	0.54	86.49	98.98	61.00	9.00
Redes Neuronales	7	0.07	89.18	98.00	81.63	5.90
Regresión Logística	14	0.14	85.70	98.95	45.00	13.94
Regresión Lineal	3	0.03	87.20	89.32	84.48	2.48
Support vector machine	2	0.02	79.00	98.00	60.00	26.87
Random Forest	8	0.08	91.31	95.00	86.20	3.08
Método bayesiano	9	0.09	81.90	95.00	67.00	10.78
Reglas de asociación	1	0.01	71.00	71.00	71.00	-
KDD	1	0.01	62.00	62.00	62.00	-
Support Vector R.	1	0.01	91.32	91.32	91.32	-
Total(T)	**99**	**1.00**				

Random Forest con una frecuencia del 8% presenta una precisión media del 91.31% con una desviación estándar del 3.08%, pero al igual que las redes neuronales adolece de no presentar una frecuencia alta, lo que deja duda acerca de la confiabilidad de estos indicadores.

En el caso de los demás modelos, por presentar una frecuencia aun menor o una desviación estándar bastante alta, no se han considerado relevantes los resultados obtenidos, por lo que será necesario tener mayor cantidad de ejemplos para valorarlos de mejor manera.

En el estudio propuesto por Kabathova y Drlik (2021), se compararon seis algoritmos de aprendizaje automático (Arboles de decisión, Redes Neuronales, Regresión Logística, Support vector machine, Random Forest, Método bayesiano), utilizando un conjunto de datos sobre la actividad de 261 estudiantes de tercer año en el curso universitario. Los modelos fueron entrenados y probados en el conjunto de datos balanceado de tres años académicos.

Se observa que los seis algoritmos mencionados anteriormente, se ven reflejados en la presente investigación, dentro de los modelos de mayor frecuencia de aplicación en problemas dedicados a la predicción de casos de abandono de estudio. Esto quiere decir que al menos el 60% de los modelos identificados en la etapa de la búsqueda de información fueron considerados como los más eficientes de acuerdo con los valores de la precisión media y de la desviación estándar.

Hay que señalar que lo planteado en el estudio de Kabathova y Drlik (2021) acoge otro tipo de metodología de investigación, y esto se evidencia claramente desde que se menciona que los modelos fueron sometidos a las fases de entrenamiento y prueba con respecto a un determinado conjunto de datos. Por otro lado, esta investigación, sigue otra línea metodológica denominada revisión sistemática, es decir, no se presenta una fase en la que se evalúan los modelos predictivos, sino que más bien, se recogen datos específicos de artículos en los que ese proceso se haya evaluado concretamente. Sin embargo, es importante aludir que independientemente de la metodología, se pretende llegar al mismo resultado, identificando el modelo más eficiente para predecir casos de abandono de estudios.

La precisión de la predicción varió entre 71% y 93%, lo que indica que, independientemente del modelo utilizado, las características elegidas en este estudio demostraron ser exitosas para predecir la continuidad o la deserción de los estudiantes a pesar de la limitación del conjunto de datos y la cantidad de características. Los árboles de decisión son 71% precisos, a diferencia del 86.49% identificado en la presente investigación como uno de los modelos con mayor precisión; este cambio drástico probablemente ocurra por la distribución balanceada de los datos, y el estado de acoplamiento de los algoritmos de acuerdo con esa distribución. Si se observa la Tabla 3, se evidencia como

Random Forest refleja un 91.31% de precisión media, y en esta ocasión en un caso experimental se obtiene un valor medianamente igual, que corresponde al 86%. De todos los modelos identificados con frecuencia de aplicación considerable, Random Forest tiene la menor desviación estándar, 3.08 respectivamente. Entonces, es válido inferir que este modelo tiene que ser considerado como un parámetro crítico de análisis para la toma decisiones en la elección de un modelo en específico.

Muy aparte de esto, el modelo Naïve Bayes y las redes neuronales obtuvieron 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 supervisados depende en gran medida de suficientes registros y que estos estén balanceados, lo que supone que, desde el punto de vista cuantitativo, el tamaño y equilibrio del conjunto de datos será otro campo determinante para valorar la eficiencia de un modelo u otro.

4. CONCLUSIONES

Los modelos de árboles de decisión son los más utilizados para la predicción de casos de abandonos de estudios, esto se debe a su gran estabilidad a la hora de obtener los resultados de la frecuencia de aplicación y de la precisión media, presentando valores bastante aceptables en comparación de otros modelos predictivos.

Mas del 90% de las ocasiones, según la literatura consultada, se implementó un modelo de aprendizaje supervisado, ya que es muy probable que no hayan tenido problemas con la definición de la clase objetivo dado la arquitectura del caso de estudio. Haciendo un análisis, la clase objetivo se puede dividir en 2 valores: cuando los estudiantes abandonan sus estudios, y el respectivo caso contrario; entonces es aceptable inferir que los autores consideraron a la variable objetivo como un problema binario y que además esta variable formaba parte del conjunto de datos con que trabajaron.

Los modelos que presentaron los valores de desviación estándar más bajos y por ende una menor variabilidad en la precisión con respecto a su media fueron: Arboles de decisión, redes neuronales y random forest, dándonos a entender que estos son modelos para considerar para la aplicación del modelo. De esta forma se puede explicar la importancia de este parámetro dado que identifica los modelos más estables y en consecuencia más eficientes para esta problemática.

REFERENCIAS

Abdulrahman, S. A., Khalifa, W., Roushdy, M., & Salem, A. B. M. (2020). Comparative study for 8 computational intelligence algorithms for human identification. Computer Science Review, 36, 100237. https://doi.org/10.1016/j.cosrev.2020.100237

Carrizo, D. & Moller, C. (2018). Estructuras metodológicas de revisiones sistemáticas de literatura en Ingeniería de Software: un estudio de mapeo sistemático. Ingeniare, 26, 45-54.

Gonzembach, J. D., Demetrio, W., & Delgado, D. (2021). Inteligencia computacional para la evaluación de las capacidades coordinativas de los estudiantes. Computational intelligence for the evaluation. 14(4), 271–287. https://publicaciones.uci.cu/index.php/serie/article/view/749/734

Guerra, L., Rivero, D., Ortiz, A., Diaz, E., & Quishpe, S. (2020). Minería de datos y uso de inteligencia computacional para la determinación de perfiles de insolvencia económica. Revista Ibérica de Sistemas e Tecnologias de Informação, E35, 48–61.

Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. Applied Sciences (Switzerland), 11(7). https://doi.org/10.3390/app11073130

Oprea, M. (2020). A general framework and guidelines for benchmarking computational intelligence algorithms applied to forecasting problems derived from an application domain-oriented survey. Applied Soft Computing Journal, 89, 106103. https://doi.org/10.1016/j.asoc.2020.106103

Pandolfi, D., Villagra, A., & Molina, D. (2017). Inteligencia computacional aplicada a la optimización multiojetivo de problemas de scheduling con restricciones. 91–94. http://sedici.unlp.edu.ar/handle/10915/61503

Telikani, A., Gandomi, A. H., & Shahbahrami, A. (2020). A survey of evolutionary computation for association rule mining. Information Sciences, 524, 318–352. https://doi.org/10.1016/j.ins.2020.02.073