https://doi.org/10.36315/2024v2end093
Education and New Developments 2024

MACHINE LEARNING PREDICTION OF ACADEMIC PERFORMANCE OF LATIN AMERICAN UNIVERSITY STUDENTS. A REVIEW

Dayana Barrera[1], Carlos Fresneda-Portillo[2], & Ana María Pacheco-Martínez[2]
[1]UNAD Colombia (Colombia)
[2]Universidad Loyola (Spain)

Abstract

Machine Learning (ML) is increasingly recognized as a powerful tool in predicting academic performance, providing essential aid to educational institutions in identifying at-risk students, facilitating timely interventions and enhancing overall student retention. This article presents a systematic review of literature over the past ten years from recognized databases with a particular focus on the prevalent ML algorithms employed in Latin American and Higher Education institutions of alike emerging countries for predicting student performance. This review reveals a significant efficacy of supervised learning models, especially Decision Trees and Neural Networks with accuracy metrics above the 80%. The review showed that the accuracy of the method depends on the quality and features of the student data available to train the model. Last, we list the most common student factors that contribute in these algorithms to predict student performance. There is no general rule to choose which student features must be included, but the literature shows that they may depend on the subject are or the specific predictive purpose of the algorithm.

Keywords: *Neural networks, decision trees, student performance, higher education, Latin America.*

1. Introduction

The term "Machine Learning" (ML) was defined by Arthur Samuel in the 1950s as *a field of study that gives machines the ability to learn about something for which they have not been explicitly programmed* (Wiederhold et al., 1990). In general, ML is considered to be a subfield of Artificial Intelligence (AI) whose purpose is analyzing algorithms to identify patterns, relationships, trends and predictions that allow a better understanding of the behavior of data in a certain phenomenon.

Artificial intelligence (AI) is being widely applied in Higher Education (HE) Institutions (HEIs) in Latin America, providing various applications to improve university services. One of these applications is, for instance, offering more personalized and effective help. Especially for those students who find themselves in difficult situations (Chen, 2011), that is, for those students who face learning difficulties, academic lags, or adverse socioeconomic factors. Providing timely student support may influence is beneficial for HEIs since it could contribute towards improving retention rates.

Another relevant application clearly linked to retention rates is the prediction of student academic success in their studies or prediction of the risk of failure in a certain module. Overall, the prediction of student success or failure has become a relevant challenge for academic institutions across the globe since student retention has direct consequences in financial and resource planning.

A possible solution to this challenge is the use of ML algorithms. In the last decade, there has been a growing literature on ML algorithms fed with educational and academic data. Therefore, ML has recently become an invaluable method for predicting student performance. These tools provide an opportunity to identify early warning signs that lecturers, coordinators and managers could use to potentially prevent certain students from dropping out.

Not only ML has shown capabilities from flagging students at risk of dropping out or failing a module, but also it has been able to identify patterns, behaviors, factors affecting this risk (Albreiki et al., 2021). Understanding which factors may have an impact on student's performance or the student's decision for dropping out, could help to the design of tailored interventions to mitigate the negative effects of this factors.

Despite the emerging literature in this topic as well as the fast development of new ML techniques, there are neither clear guidelines that suggest which ML technique is more suitable for this purpose nor information about which data should be collected to train the ML algorithms. To overcome this issue, we have developed a systematic review of the literature published in this topic with the purpose of answering the following questions research questions:

RQ1 - What are the most common ML techniques for predicting student drop-out / student performance in HEIs within Colombia, Latin America and similar emerging countries?

RQ2 - Which of these ML techniques happen to be more accurate?

RQ3 - What are the dominant factors affecting student drop-out / student performance according to these ML techniques found in RQ1 and RQ2?

2. Methods

In this section, the procedure followed for the systematic literature review is detailed. See also Figure 1.

In the first step of the literature search, manuscript written in Spanish with ML techniques applied within Colmbian HEIs were prioritized. As a result, 11 valid works were obtained; therefore, the search was expanded to other Latin American countries, but not enough references were found in Spanish. Therefore, articles written in English and studies conducted in other developing countries were included in the search criteria. To decide which countries are to be included, the GNI per capita and development and human resources indices were taken into account from WorldData.info. After widening the criterion of the procedence of the study several additional countries were considered. In particular, studies from: Saudi Arabia, Iran, and Iraq from the Middle East; India, Indonesia, Pakistan, Taiwan and Malaysia from South East Asia; and from Bulgaria were included in the review.

A total of 209 articles were retrieved from the following databases: Web of Science (103 articles), Google Scholar (63 articles), Dialnet (28 articles) and SciELO (15 articles). Although the Scopus database was included in the search process, it did not yield any additional articles beyond those already identified in the previous databases. These articles were found by consistently using the search question: ("Student performance" OR "Drop out") AND ("prediction") AND ("Machine Learning"), filtering by Colombia or the Latin American region. The search was conducted in both languages, English and Spanish.

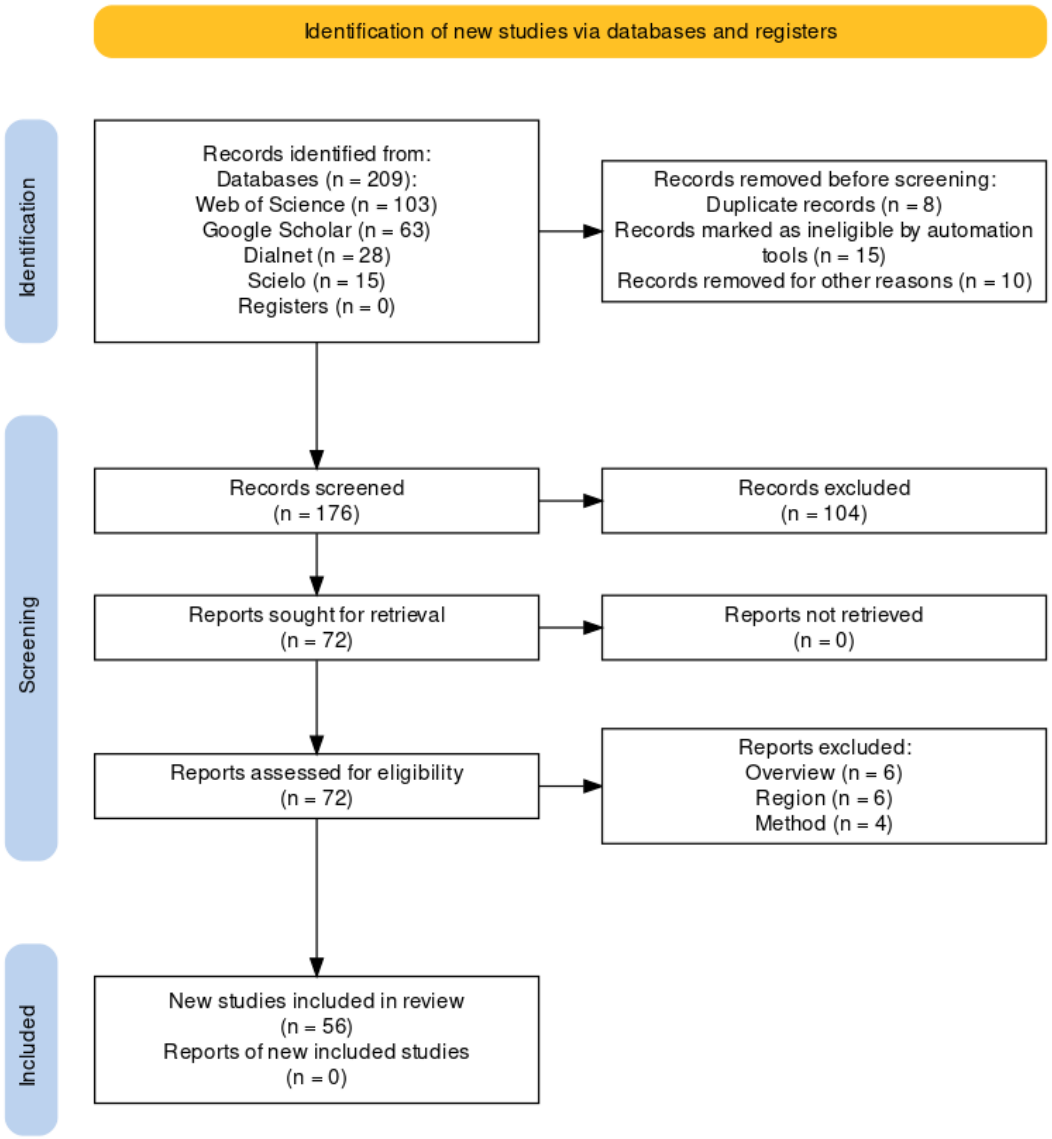
Several filters were applied to the 209 articles found. The first filter consisted of removing 33 articles that appeared duplicated. Then, a second filter of open access was applied and thus, 104 articles were excluded. A rationale for including this filter is a future comparison between algorithms published in open access against those published in traditional form. Finally, a third filter: *the study contributes towards the objectives of this review, i.e., region and methods based on ML* was applied. As a result, 16 articles were excluded. Therefore, a total of 56 articles were included in the review (see Figure 1).

The inclusion and exclusion criteria used to select articles for this review are detailed in Table 1.

Table 1. Inclusion and exclusion criteria.

Criterion	Inclusion	Exclusion
Topic	ML techniques	Qualitative methods
Source	Journal papers	Other sources
Publication year	2013-2023	Other
Language	Español, Inglés	Other languages

Figure 1. PRISMA flow diagram.



3. Key concepts of machine learning for education professionals

In this section, we include the definition, an explanation, and an example of application of the key three concepts of ML that will be required to understand the results of the next section (Mueller & Massaron, 2021). First, we say that a ML algorithm is supervised if it involves training a model with a given dataset with a certain set of input variables and their corresponding associated output. The model learns to map inputs to outputs based on this data and can then predict the output for new, unseen data. It is said to be *supervised* because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. Two important examples of supervised techniques are: decision trees and neural networks

On the one hand, decision trees are similar to tree diagrams for decision making. At each node there is a condition which divides the data into two different subsets. The condition is chosen by the algorithm to best separate the data. Then, at each subset, the tree continues to branch until there are no data features which separate the data, i.e., the subset is homogeneous. An example of application is the decision-making process for a bank to grant a loan. The bank will use customer features such as credit score, annual income, employment status, etc. The decision tree will create branches for each of these features, categorizing the applications into different risk classes (e.g., high risk, medium risk, low risk), helping the bank decide whom to offer a loan.

On the other hand, neural networks are inspired by the structure of the human brain and consist of layers of interconnected nodes or neurons. Each connection between nodes has an associated weight, which is adjusted during training. Each neuron works as a logistic-type regression model. What happens in the subsequent layer, depends on the outcomes of the previous layer. Therefore, one can think of a neural network as a complex set of many multiple regression models. One application of neural networks is object recognition within photographs.

4. Results

First, it is worth noting the increasing interest in this topic shown in the rise on the number of publications in recent years. Out of these 56 studies, 38 date from 2020 to 2023. Second, regarding the HEI country of the study, 80% of the studies included in the review are from Latin American countries.

With regards to RQ1, 78.6% of the studies apply supervised ML models. This is because most HEIs own data about students' performance in previous years. Hence, they can exploit their data to train a ML model and then test it with data of subsequent years. After several testing iterations, it is expected that the model is validated and thus, can be applied to accurately predict the outcome of interest.

Moving on to RQ2, the predominant ML algorithms applied within these models use Decision Trees (DT), Neural Networks (NN) and Random Forest (RF). DT appear in 31.6% of the studies while NN and RF appear with a frequency of 23.7% and 10.5% respectively. According to (Natek and Zwilling, 2014) universities may not hold sufficient data to create a sufficiently robust DT based ML algorithm to accurately predict student performance in general. However, when the algorithm is restricted to predict success within a certain undergraduate program it can be successful depending on the amount of data from previous years available to train the algorithm. For example, within Engineering, a DT algorithm was able to accurately predict student performance in 96.5% of the cases (Buenaño-Fernández et al, 2019). Another example applied in a Foundation Year in Education, the DT predicted with an accuracy of 91.67%, see (Díaz-Landa et al., 2021). Nevertheless, some authors defend that the accuracy of the prediction depends on the student ability. For instance, it is easier to predict students at risk of dropping out than excellent students, see (Kabakchieva, 2013). On the other hand, NNs show slightly lower degree of accuracy. For instance, the works (Su et al., 2022) and (Jishan et al., 2015) report an accuracy of 88% -86% respectively.

Last but not least, there is a wide range of variables involved in the prediction of student performance. The variables affecting the prediction depend on what exactly we do intend to predict. For example, a manager or head of school might be more interested in whether a student will drop-out of their program whereas a lecture will rather be more interested in whether a student will fail the module. Out of 522 attributes identified in this review, the factors more recurrent are, in decreasing frequency order: gender, age, first year of undergraduate study, grades, family, parents, finances, demographics, number of passed modules and subject. The effect of this factors has been previously studied in the literature, see e.g. (Ramirez and Grandon, 2018) or (Castrillón et al, 2020) and more references therein.

5. Conclusions and further work

In this study, we systematically analyzed 56 papers related to prediction of student performance in a Latin American and emerging country context. The most widely applied and accurate machine learning model for this purpose is the use of decision trees and neural networks with an accuracy of 90%. It is recommended to develop a model for subject area instead of a general model.

References

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of students' performance prediction using machine learning techniques. *Education Sciences, 11*(9), 552. https://doi.org/10.3390/educsci11090552

Buenaño-Fernández, D., Gil, D., & Luján-Mora, S. (2019). Application of machine learning in predicting performance for computer engineering students: A case study. *Sustainability, 11*(10), 2833. https://www.mdpi.com/2071-1050/11/10/2833/htm

Castrillón, O. D., Sarache, W., & Ruiz-Herrera, S. (2020). Prediction of academic performance using artificial intelligence techniques. *Formación Universitaria, 13*(1), 93-102. http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-50062020000100093&lng=en&nrm=iso&tlng=en

Chen, L. H. (2011). Enhancement of student learning performance using personalized diagnosis and remedial learning system. *Computers and Education, 56*(1), 289-299. https://doi.org/10.1016/j.compedu.2010.07.015

Chen, F., & Cui, Y. (2020). Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics, 7*(2), 1-17. https://learning-analytics.info/index.php/JLA/article/view/6777

Díaz-Landa, B., Meleán-Romero, R., & Marín-Rodriguez, W. (2021). Rendimiento académico de estudiantes en educación superior: Predicciones de factores influyentes a partir de árboles de decisión. *Telos: Revista de Estudios Interdisciplinarios en Ciencias Sociales, 23*(3), 616-639. https://doi.org/10.36390/TELOS233.08

Flores, F. A. I., Sanchez, D. L. C., Urbina, R. O. E., Soto, J. A. D., & Medrano, S. E. V. (2021). Diseño e implementación de una red neuronal artificial para predecir el rendimiento académico en estudiantes de Ingeniería Civil de la UNIFSLB. *Veritas et Scientia, 10*(1), 107-117. https://revistas.upt.edu.pe/ojs/index.php/vestsc/article/view/464/397

Jishan, S. T., Rashu, R. I., Haque, N., & Rahman, R. M. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics, 2*(1), 1-25. https://doi.org/10.1186/S40165-014-0010-2

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies, 13*(1), 61-72. https://doi.org/10.2478/cait-2013-0006

Mueller, J. P., & Massaron, L. (2021). *Machine learning for dummies*. John Wiley & Sons.

Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications, 41*(14), 6400-6407.

Su, Y. S., Lin, Y. D., & Liu, T. Q. (2022). Applying machine learning technologies to explore students' learning features and performance prediction. *Frontiers in Neuroscience, 16*, 2211. https://doi.org/10.3389/FNINS.2022.1018005/BIBTEX

Suarez Barón, M., Tinjaca Cristancho, C., & González Sanabria, J. (2020). Analítica de datos aplicada al estudio de deserción estudiantil en la Universidad Pedagógica y Tecnológica de Colombia - UPTC. *Aglala, 11*(1), 284-301. https://dialnet.unirioja.es/servlet/articulo?codigo=8458719&info=resumen&idioma=ENG

Wiederhold, G., McCarthy, J., & Feigenbaum, E. (1990). Arthur Samuel: Pioneer in machine learning. *Communications of the ACM, 33*(11).