Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review

Daniel Andrade-Girón¹, Juana Sandivar-Rosas ², William Marín-Rodriguez¹*, Edgar Susanibar-Ramírez³, Eliseo Toro-Dextre¹, Jose Ausejo-Sanchez¹, Henry Villarreal-Torres³, Julio Angeles-Morales³

¹ Universidad Nacional José Faustino Sánchez Carrión. Huacho, Perú
² National University of San Marcos, Perú.
³ Universidad San Pedro. Chimbote, Perú.

Abstract

Student dropout is one of the most complex challenges facing the education system worldwide. In order to evaluate the success of Machine Learning and Deep Learning algorithms in predicting student dropout, a systematic review was conducted. The search was carried out in several electronic bibliographic databases, including Scopus, IEEE, and Web of Science, covering up to June 2023, having 246 articles as search reports. Exclusion criteria, such as review articles, editorials, letters, and comments, were established. The final review included 23 studies in which performance metrics such as accuracy/precision, sensitivity/recall, specificity, and area under the curve (AUC) were evaluated. In addition, aspects related to study modality, training, testing strategy, cross-validation, and confounding matrix were considered. The review results revealed that the most used Machine Learning algorithm was Random Forest, present in 21.73% of the studies; this algorithm obtained an accuracy of 99% in the prediction of student dropout, higher than all the algorithms used in the total number of studies reviewed.

Keywords: prediction, student attrition, machine learning, deep learning.

Received on 11 December 2022, accepted on 07 July 2023, published on 18 July 2023

Copyright © 2023 Girón et al., licensed to EAI. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/eetsis.3586

¹Corresponding author. Email: wmarin@unjfsc.edu.pe

1. Introduction

Student dropout is widely recognized worldwide as one of the most complex challenges facing the education system ¹,², and this phenomenon has experienced a significant increase during the COVID-19 pandemic ³. This issue entails economic, social, and educational consequences for the stakeholders in the global education system, ranging from the psychological impact on students to the management challenges faced by government entities ⁴,⁵. To address the problem, predicting and managing early signs of student dropout is relevant ⁶–⁹. This will enable educational institutions to act promptly, implementing preventive and proactive measures to address the issue and reduce the dropout rate ¹⁰. Various governments have designed and implemented early warning systems for school dropouts to effectively tackle this problem ¹¹–¹³.

An alternative of great relevance is using Machine Learning and Deep Learning algorithms ¹⁴. These models predict dropout and provide early alerts to relevant authorities, enabling them to take alternative measures targeted at at-risk students ¹⁵.

Each Machine Learning and Deep Learning model is intrinsically linked to the underlying algorithm, optimized hyperparameters, the training and test datasets used ¹⁶, as well as the variables and data behavior, different performance metrics ¹⁷. As a result, multiple alternatives are
observed, offering different results in each specific application. Consequently, Machine Learning and Deep Learning approaches have been subject to criticism due to their use of a "black box" methodology in predicting student dropout, which results in a lack of proper interpretability of the model for humans. Therefore, conducting a comprehensive systematic review study on the application of Machine Learning and Deep Learning in predicting student dropout is imperative. This study aims to identify algorithms that have demonstrated better predictive capabilities and the different variants of each model.

A thorough search has been conducted in the major databases of systematic review studies related to student dropout. However, research specifically oriented toward our purpose has yet to be found. Therefore, our main objective is to fill this knowledge gap and answer which Machine Learning and Deep Learning algorithms perform best in predicting student dropout.

2. Methods

The present research has been developed using the systematic review methodology based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The following phases were followed in developing the systematic review: Firstly, the research question guiding the study was formulated. Then, a research protocol was developed describing the design of the systematic review, including the criteria for study selection, the databases used for the search, the search strategies, and the methods of data extraction and analysis. Inclusion and exclusion criteria for the studies were also established.

Subsequently, an exhaustive search of the scientific literature was conducted in different databases, using the defined search terms and applying the inclusion and exclusion criteria to select relevant studies. The titles and abstracts of the articles identified in the search were reviewed, selecting those that met the established inclusion criteria. A full reading of the selected studies was then conducted to verify their compliance with the inclusion and exclusion criteria. The relevant data from the selected studies were extracted and organized in a database.

Finally, the results were interpreted, and the findings were synthesized, identifying possible limitations. In the last phase, a detailed report of the systematic review was written, including a complete description of the methodology used, the results obtained, and the conclusions reached.

Search Strategy
To conduct this systematic review, an exhaustive search was performed in specialized databases to find relevant information for our research. Table 1 presents a detailed description of the search strategy used.

<table>
<thead>
<tr>
<th>Database</th>
<th>Search syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scopus</td>
<td>TITLE-ABS-KEY (&quot;Machine learning&quot;) AND TITLE-ABS-KEY (&quot;student dropout&quot; OR &quot;Student desertion&quot;)</td>
</tr>
<tr>
<td>IEEE</td>
<td>(&quot;All Metadata&quot;:&quot;Machine Learning&quot;) AND (&quot;All Metadata&quot;:&quot;student dropout&quot;) OR (&quot;All Metadata&quot;:&quot;student desertion&quot;)</td>
</tr>
<tr>
<td>Xplore</td>
<td>(&quot;All Metadata&quot;:&quot;student dropout&quot;) OR (&quot;All Metadata&quot;:&quot;student desertion&quot;)</td>
</tr>
<tr>
<td>Library Web</td>
<td>(TI=(performance OR achievement)) AND</td>
</tr>
<tr>
<td>Science</td>
<td>ALL=(collaboration)) AND ALL=(programming)</td>
</tr>
</tbody>
</table>

Inclusion and exclusion criteria
The inclusion and exclusion criteria in this scientific research refer to the standards and rules established to determine which studies or articles will be considered in the systematic review and which will be excluded. These criteria are based on the research objectives and questions being addressed.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Inclusion</th>
<th>Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>Basic and higher education students</td>
<td>Postgraduate students</td>
</tr>
<tr>
<td>Phenomenon of interest</td>
<td>Student desertion</td>
<td></td>
</tr>
<tr>
<td>Time Period</td>
<td>Studies: from 2000 to 2023</td>
<td>Studies outside this time interval</td>
</tr>
<tr>
<td>Languages</td>
<td>English</td>
<td>Language other than English</td>
</tr>
<tr>
<td>Focus of the study</td>
<td>Quantitative approach</td>
<td>Qualitative approach</td>
</tr>
</tbody>
</table>

Sample Selection Process
After applying the inclusion and exclusion criteria, the sample was restricted to analyzing only those articles that provided information relevant to the objective set. The attached flow chart shows that 246 articles were initially identified in the three databases. After eliminating duplicate articles and applying the inclusion and exclusion criteria, 43 articles were obtained. From this selection, additional exclusions were made for various reasons. In the end, a total of 23 articles were included in the analysis (Figure 1).
3. Results and Discussion

Table 3 presents the most relevant characteristics for the systematic review. The following attributes have been considered: author, country, sample, number of variables, training strategy, cross-validation, modality, Machine Learning and Deep Learning algorithm, performance metric, best-performing algorithm, and results (accuracy, sensitivity/recall, F1 score).

The results of this study show the distribution of selected articles according to their country of origin. Most articles (21%) were published in China, while 17.39% originated from the United States. Additionally, 8.69% of the selected articles came from Korea, India, and Spain. Other countries contributing to the sample included Turkey, Hungary, Germany, Malaysia, Chile, Ecuador, Slovakia, and the Netherlands, representing 4.34% of the selected articles. These findings suggest that student dropout is a relevant research topic in various parts of the world. However, the selection of the used database may have influenced this distribution.

According to the results obtained, it was observed that 100% of the studies included in the systematic review mostly employed supervised Machine Learning algorithms for classification. The total sum of samples was 1,912,653, with a mean of 57,959. Furthermore, the total number of variables was 373, with a mean of 16.21.

Regarding the training and test sets, the following patterns were found: 30.43% of the studies used 70% of the sample for training, while 26.08% used 30% for testing. Additionally, 17.39% of the studies allocated 80% of the sample for training and 20% for testing. Likewise, 8.69% of the studies used 60% for training and 40% for testing, while 4.34% employed 90% for training and 10% for testing. Finally, only 4.34% of the studies focused on validation, while a similar percentage (39.13%) did not report evidence for both training and testing.

Regarding the cross-validation strategy, it was observed that 60.86% of the studies used the 10-fold cross-validation method, 13.04% opted for the 5-fold method, and 4.34% used the 9-fold method, while 17.39% of the studies did not report the value of k-fold.

Regarding the modality of the studies, it was found that 69.56% corresponded to the in-person modality, while 30.43% belonged to the virtual modality. Additionally, it was observed that 56.52% of the studies used ROC-AUC validation tests, while 43.47% did not report using such tests.

When analyzing the Machine Learning and Deep Learning algorithms used in the studies, the following results were found: 39.13% used neural networks corresponding to Deep Learning, 56.52% used decision trees, 39.13% used logistic regression, 30.43% used support vector machines, 47.82% used Random Forest, 13.04% used Gradient Boosted Tree, 17.39% used Naïve Bayes, 8.69% used Generalized Linear Model, 13.82% used k-nearest neighbors, 4.34% used Ada boost, 8.69% used XG Boost, 4.34% used Cat Boost, 4.34% used Free-Forward, 4.34% used Stacking Ensemble, 4.34% used Bayesian networks, and 4.34% used Ripper.

Regarding the Machine Learning and Deep Learning algorithms that achieved the best performance in each study, the following results were found in about 100% of the studies:

- 21.73% reported Random Forest as the best-performing algorithm.
- 13.04% reported that Logistic Regression and decision trees achieved the best performance in equal proportion.
- 8.69% reported Gradient Boosted Tree as the best-performing algorithm.
- 4.34% reported that Stacking Ensemble, Boosted Decision, SVM, CART Model cost3, K-NN, and CBN were the best-performing algorithms in each study.

Regarding performance, it was observed that the Random Forest algorithm achieved an accuracy of 99%, representing the highest performance obtained in the research works. This result finds theoretical support in the literature, as Random Forest has demonstrated superior performance in most studies.
Table 3. Most relevant characteristics for the systematic review.

<table>
<thead>
<tr>
<th>Author</th>
<th>Country</th>
<th>Sample</th>
<th>No. of variables</th>
<th>Training strategy</th>
<th>Cross-validation</th>
<th>Modality</th>
<th>Algorithm Used for comparison</th>
<th>Performance metrics</th>
<th>Chosen algorithm</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Kiss, Maldonado, &amp; Segall, 2022)</td>
<td>USA</td>
<td>21,079</td>
<td>07</td>
<td>07</td>
<td>17%</td>
<td>On-site</td>
<td>Neuronal Networks (NN), Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM)</td>
<td>SI</td>
<td>SI</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>(Nagy &amp; Molontay, 2018)</td>
<td>Hungary</td>
<td>15,825</td>
<td>36</td>
<td>10-fold</td>
<td>On-site</td>
<td>yes</td>
<td>Neuronal network (NN), Decision Tree (DT), Random Forest (RF), Gradient Boosted Tree, Logistic Regression (LR), Naïve Bayes (NB), Generalized Linear Model (LightGBM), K-NN, Adaptive Boosting (AdaBoost)</td>
<td>0.769</td>
<td>Gradient Boosted Tree (GBDT)</td>
<td>76.6 %</td>
</tr>
<tr>
<td>(Rodríguez, Villanueva, Dombrowskaia, &amp; Velenzuela, 2023)</td>
<td>Chile</td>
<td>691,748</td>
<td>26</td>
<td>80%</td>
<td>On-site</td>
<td>no</td>
<td>Decision tree (DT), XGBoost, LightGBM, CatBoost</td>
<td>no</td>
<td>LightGBM</td>
<td>93%</td>
</tr>
<tr>
<td>(Sandoval-Palis, Naranjo, Vidal, &amp; GilarrCorbi, 2020)</td>
<td>Ecuador</td>
<td>2,097</td>
<td>4</td>
<td>70%</td>
<td>30%</td>
<td>On-site</td>
<td>Neuronal network (NN), Logistic Regression (LR)</td>
<td>yes</td>
<td>Neuronal network (NN)</td>
<td>77%</td>
</tr>
<tr>
<td>Study</td>
<td>Country</td>
<td>N</td>
<td>Percentage Dropout</td>
<td>Holdout Type</td>
<td>Methodologies</td>
<td>Ensemble</td>
<td>Accuracy (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>----------</td>
<td>-------</td>
<td>--------------------</td>
<td>--------------</td>
<td>--------------------------------------------------</td>
<td>----------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Niyogisubizo, Liao, Nziyumv a, Murwana shyaka, &amp; Nshimyumukiza (2022)</td>
<td>China</td>
<td>261</td>
<td>12</td>
<td>10-fold On-site</td>
<td>Random forest(RF), XGBoost, Gradient Boosting (GB), Feedforward Neural Networks(FN N), Stacking ensemble</td>
<td>yes</td>
<td>93%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Tan &amp; Shao, 2015)</td>
<td>China</td>
<td>623</td>
<td>26</td>
<td>10-fold Online</td>
<td>Neuronal network (NN), Decision Tree(DT), Bayesian networks (RB)</td>
<td>no</td>
<td>94.6%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Dass, Gary, &amp; Cunningham, 2021)</td>
<td>USA</td>
<td>317</td>
<td>5</td>
<td>Kappa Cohen</td>
<td>Random Forest (RF)</td>
<td>yes</td>
<td>87.5%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Kemper, Vorhoff, &amp; Wigger, 2020)</td>
<td>Germany</td>
<td>620</td>
<td>16</td>
<td>10-fold On-site</td>
<td>Decision Tree(DT), Logistic Regression (LR)</td>
<td>no</td>
<td>95%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Aulck, Velagapudi, Blumenstock, &amp; West, 2016)</td>
<td>USA</td>
<td>32.5</td>
<td>7</td>
<td>Logistic Regression (LR), Random Forest (RF), K-NN</td>
<td>yes</td>
<td>66.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Yaacob, Sobri, Nasir, Norshahidi, &amp; Husin, 2020)</td>
<td>Malaysia</td>
<td>64</td>
<td>27</td>
<td>10-fold On-site</td>
<td>Logistic Regression (LR), K-NN, Random Forest (RF), Neuronal network (RN), Decision Tree (DT)</td>
<td>1</td>
<td>90%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Kabathova &amp; Drlik, 2021)</td>
<td>Slovakia</td>
<td>261</td>
<td>5</td>
<td>10-fold On-site</td>
<td>Naive Bayes (NB), Random Forest (RF), Randomized Neural Network (NN), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT)</td>
<td>0.96</td>
<td>93%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Lee &amp; Chung, 2019)</td>
<td>South Korea</td>
<td>165, 715</td>
<td>15, 80</td>
<td>10-fold On-site</td>
<td>Random Forest (RF), Boosted decision tree (DBT)</td>
<td>yes</td>
<td>98%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Country</td>
<td>Methodology</td>
<td>nSMOTE, Random Forest (RF), Boosted decision tree (DBT)</td>
<td>Accuracy</td>
<td>Precision</td>
<td>Recall</td>
<td>F1-score</td>
<td>Support Vector Machines (SVM)</td>
<td>Decision Tree (DT)</td>
<td>Random Forest (RF)</td>
<td>Random Forest (RF)</td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------</td>
<td>----------</td>
<td>-----------</td>
<td>--------</td>
<td>----------</td>
<td>--------------------------------</td>
<td>------------------</td>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>South Korea</td>
<td>165, 715, 12, 80, 20, 10-fold On-site Random Forest (RF) yes yes</td>
<td>Random Forest (RF) 95% 95% 85% 89%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>India</td>
<td>650, 000, 16, 9-fold Online Decision Tree (DT), SVM, Naïve Bayes (NB), 1 1 Random Forest (RF) 99% 99% 99%</td>
<td>Random Forest (RF) 99% 99% 99% 99%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>200, 00, 7, 120, 542, 803, 62, 5-fold Online LR, SVM, RF, GBDT si si GBDT 89%</td>
<td>Random Forest (RF) 99% 99% 99% 99%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td>160, 66, 39, no, no, 10-fold On-site Decision Tree (DT), NN, SVM, LR no no SVM 81% 87% 77% 82%</td>
<td>Random Forest (RF) 99% 99% 99% 99%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>648, no, no, 10-fold On-site CART, Bayes Net, Logit, Rippert (JRip), Random Forest (RF) 86% 84%</td>
<td>Random Forest (RF) 86% 84%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>105, 5, 15, 10-fold On-site CART, C4.5, RB, Random Forest (RF), SVM Random Forest (RF) 86% 84%</td>
<td>Random Forest (RF) 86% 84%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkey</td>
<td>189, 9, 70, 30, 10-fold On-site J48 (DT), MLP (NN) no no MLP (NN) 96% 96% 97% 97%</td>
<td>Random Forest (RF) 86% 84%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>India</td>
<td>432, 10, si, si, 10-fold On-site ID3, C4.5, ADT (DT) no no C4.5 (DT) 74% 70% 96% 81%</td>
<td>Random Forest (RF) 86% 84%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>28, si, si, 5-fold Online KNN, RBF, SVM, (combination of multiple classifiers) CBN no no CBN 90% 95% 79%</td>
<td>Random Forest (RF) 86% 84%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>623, 75, 26, 70, 30, Online Neuronal network (NN), Decision Tree (DT) and Bayesian no no NN 93.9 7% 98.8 5% 94.6 3% 95%</td>
<td>Random Forest (RF) 86% 84%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In the articles' analysis, 16 algorithms applied in Machine Learning and Deep Learning have been identified. Among these algorithms, it has been observed that RandomForest has exhibited the best performance, achieving an accuracy of 99% (Table 03). Next, we will discuss the theoretical rationale behind why RandomForest has outperformed other algorithms 50–54.

The application of Machine Learning and Deep Learning algorithms poses two main challenges. Firstly, it is essential to determine the optimal algorithm, which is a complex task given multiple candidate systems that meet the established criteria 55. This problem becomes particularly challenging when the learning algorithm has a propensity for diverse local optima and insufficient training data availability 56. Secondly, by discarding less successful models, there is a risk of losing potentially valuable information 56,57.

RandomForest is a learning algorithm based on creating an ensemble of decision trees and combining their results to obtain a final prediction 58. Each tree in the ensemble is constructed independently using the technique known as "bagging" 58, which involves taking random samples with replacement from the original training dataset and building a decision tree from each of these samples 59. A theoretical justification for the superior performance of RandomForest lies in its nature as a Machine Learning ensemble 60–65, which are techniques that combine multiple individual models to improve predictive capability and system robustness 66. It is characterized by creating a set of decision trees, each representing an individual model in the ensemble 66,67–72. Each tree is constructed using a random sample with replacement from the original training dataset and a random selection of features at each node 73.

Another key aspect supporting the advantage of RandomForest is its ability to address local optima challenges. Some algorithms, such as decision trees, can generate highly non-convex cost functions, which can cause the methods used to solve them to become trapped in local optima 79. Combining different hypotheses through different approaches in each of them increases the probability of approximating the true hypothesis more accurately 74–78. This is because different solutions are explored, reducing the reliance on a single local optimum 79.

Indeed, the RandomForest algorithm has demonstrated superior performance to other algorithms in Machine Learning and Deep Learning when applied to predicting student dropout. Its ability to address local optima and overfitting challenges and leverage diversity and independence among the trees makes it a suitable choice 80. The ensemble approach of machine learning, of which RandomForest is an example, has been shown to be beneficial in combining multiple models to improve predictive capability and system robustness, reduce the risk of selecting an incorrect hypothesis, and expand the hypothesis space to more effectively approximate the target function.

2. Conclusions

This systematic review study has provided an overview of predicting student dropout using Machine Learning and Deep Learning techniques. The most promising algorithms and their variants in terms of predictive capability were identified. Timely prediction of student dropout has significant potential to improve educational management and contribute to achieving quality standards in the educational field.

After analyzing 23 scientific articles, the application of 16 different Machine Learning and Deep Learning algorithms was highlighted. The most utilized algorithm in these studies was RandomForest, representing approximately 21.73% of the total. Additionally, RandomForest demonstrated outstanding performance, achieving an impressive accuracy of 99%.

A key advantage of the RandomForest model, based on an ensemble of Machine Learning algorithms, lies in its ability to overcome local optima and overfitting issues. However, it is important to note that more variables related to student dropout and further research using large volumes of data are required to obtain more robust and generalizable results.

Overall, this study highlights the potential of Machine Learning and Deep Learning techniques in addressing the challenge of student dropout. The findings and recommendations presented in this article are expected to serve as a starting point for future research and practical applications in the educational field, aiming to improve student retention and academic success.

References

3. Del Savio AA, Galantini K, Pachas A. Exploring the relationship between mental health-related problems and undergraduate student dropout: A
Predicting University Dropout through


Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review


