EAI Endorsed Transactions
on Scalable Information Systems
Review Article EAI.EU

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

Daniel Andrade-Girón[1], Juana Sandivar-Rosas [2], William Marín-Rodriguez[1,*], Edgar Susanibar-Ramirez[1], Eliseo Toro-Dextre[1], Jose Ausejo-Sanchez[1], Henry Villarreal-Torres[3], Julio Angeles-Morales[3]

[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

doi: 10.4108/eetsis.3586

*Corresponding 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 [13,18]. Consequently, Machine Learning and Deep Learning approaches [19] 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 [20]. Therefore, conducting a comprehensive systematic review study on the application of Machine Learning and Deep Learning in predicting student dropout is imperative [21–24]. 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 [25] based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [26,27].

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 [28–30].

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 1. Search strategy for each database

Database	Search syntax
Scopus	TITLE-ABS-KEY ("Machine learning") AND TITLE-ABS-KEY ("student dropout" OR "Student desertion")
IEEE Xplore Digital Library	("All Metadata":"Machine Learning") AND ("All Metadata":"student dropout") OR ("All Metadata":"student desertion")
Web Science	((TI=(performance OR achievement)) AND ALL=(collaboration)) AND ALL=(programming)

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 2. Inclusion and exclusion criteria

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

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).

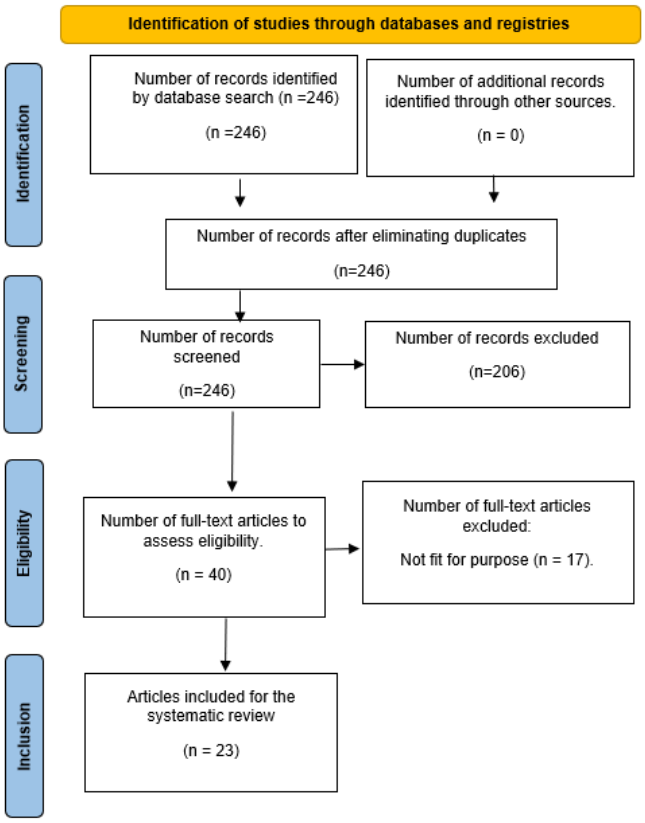

Figure 1: Flowchart of the search and selection method for the systematic review references.

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 Ensamble, 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 Ensamble, 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.

Author	Country	Sample	N° of variable	Training strategy		Cross-validation	Modality	Algorithm Used for comparison	Performance metrics		Chosen algorithm	Results			
				Train	Test				ROC	AUC		Accu	pres	ecall	F1-scor
(Kiss, Maldonado, & Segall, 2022) [31]	USA	21,079	07	17 %	15 % 15 %		On-site	Neuronal Networks (NN), Decision Tree (DT), Logistic Regression (LR). Support Vector Machine (SVM)	SI	SI	Logistic Regression	84,8 %		93.8 %	
(Nagy & Molontay, 2018) [32]	Hungary	15,825	36			10-fold	On-site	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 Boo(AdaBoost)	yes	0.769	Gradient BoostedTree (GBDT)	76.6 %	70.2 %	75%	72 %
(Rodríguez, Villanueva, Dombrovskaia, & Velenzuela, 2023) [33]	Chile	691,748	26	80 %	20 %		On-site	Decision tree(DT), XGBoost, LightGBM, CatBoost	no	no	LightGBM	93%	17%	83%	29 %
(Sandoval-Palis, Naranjo, Vidal, & Gilar-Corbi, 2020) [34]	Ecuador	2,097	4	70 %	30 %	10-fold	On-site	Neuronal network (NN), Logistic Regression (LR)	yes	yes	Neuronal network (NN)	77%			

(Niyogisubizo, Liao, Nziyumva, Murwanashyaka, & Nshimyumukiza, 2022) [2]	China	261	12	80 %	20 %	10-fold	On-site	Random forest(RF), XGBoost, Gradient Boosting (GB), Feed-forward Neural, Networks(FNN), Stacking ensemble	yes	yes	Stacking ensemble	93%	93%	93%	92 %
(Tan & Shao, 2015) [35]	China	62375	26	70 %	30 %		Online	Neuronal network (NN), Decision Tree(DT), Bayesian networks (RB)	no	no	Decision Tree(DT)	94.63%	65%	82%	72 %
(Dass, Gary, & Cunningham, 2021) [36]	USA	3172	5			Kappa cohen	Online	Random Forest (RF)	Yes	94.5	Random Forest (RF)	87.5 %	88%	87.5 %	87.5%
(Kemper, Vorhoff, & Wigger, 2020) [37]	Germany	620	16	90	10	10-fold	On-site	Decision Tree(DT), Logistic Regressión(LR)	no	no	Decision Tree (DT)	95%	94%	98%	84 %
(Aulck, Velagapudi, Blumenstock, & West, 2016) [38]	USA	32,500	7	70 %	30 %	10-fold	On-site	Logistic Regression (LR), Random Forest(RF),K-NN	Yes	Yes	Logistic Regression	66.59%			
(Yaacob, Sobri, Nasir, Norshahidi, & Husin, 2020) [39]	Malaysia	64	27	60 %	40 %	10-fold	On-site	Logistic Regression (LR), K-NN, Random Forest (RF), Neuronal network (RN), Decision Tree (DT)	1	87 %	Logistic Regression	90%	100 %	100 %	100 %
(Kabathova & Drlik, 2021) [18]	Slovakia	261	5			10-fold	On-site	Naïve Bayes (NB) ,Random Forest(RF),Red Neuronal (NN), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT)		0.96	RandomForest(RF)	93%	86%	96%	91 %
(Lee & Chung, 2019) [40]	South Korea	165,715	15	80	20	10-fold	On-site	Random Forest(RF),Boosted decision ,ree(DBT),Co	yes	yes	Boosted decision tree(DBT)	98%	98%	89%	93 %

								n SMOTE,Random Forest(RF),Boosted decision tree(DBT)							
(Chung & Lee, 2019) [13]	South Korea	165,715	12	80	20	10-fold	On-site	Random Forest(RF)	yes	yes	Random Forest(RF)	95%	95%	85%	89%
(Kashyap & Nayak, 2018) [41]	India	650000	16			9-fold	Online	Decision Tree(DT),SVM, Naïve Bayes(NB),	1	1	Random Forest(RF)	99%	99%	99%	99%
(Liang, Li, & Zheng, 2016) [42]	China	20000	7	120542	80362	5-fold	Online	LR,SVM,RF, GBDT	si	si	GBDT	89%			
(Delen, 2010) [43]	USA	16066	39	no	no	10-fold	On-site	Decision Tree(DT),NN, SVM,LR	no	no	SVM	81%	87%	77%	82%
(Dekker, Pechenizkiy, & Vleeshouwers, 2009) [44]	Netherlands	648		no	no	10-fold	On-site	CART,Bayes Net,Logit, Ripper(JRip), RF	no	no	CART Model cost 3	79%	80%	78%	79%
(Rodriguez-Muñiz, Bernardo, Esteban, & Díaz, 2019) [45]	Spain	1055	15			10-fold	On-site	CART, C4.5, RB, Radom Forest (RF), SVM			RandomForest(RF)	86%		84%	
(Lázaro, Callejas, & Griol, 2020) [46]	Spain	456	25	70	30	5-fold	On-site	J48(DT), MLP(NN)	no	no	MLP(NN)	96%	96%	97%	97%
(Yukseltu rk, Ozekes, & Turel, 2014) [47]	Turkey	189	9	70	30	10-fold	On-site	K-NN, Decision Tree (DT), Naïve Bayes (NB),Red Neuronal (NN)	yes	yes	3-NN	80%		87%	
(Yadav, Bharadwaj, & Pal, 2012) [48]	India	432	10	si	si	10-fold	On-site	ID3, C4.5, ADT (DT)	no	no	C4.5(DT)	74%	70%	96%	81%
(Dewan, Lin, & Wen, 2015) [49]	China		28	si	si	5-fold	Online	KNN, RBF,SVM,(combination of multiple classifiers) CBN	no	no	CBN	no	90%	95%	79%
(Tan & Shao, 2015) [35]	China	62375	26	70	30		Online	Neuronal network (NN), Decision Tree (DT) and Bayesian	no	no	NN	93.97%	98.85%	94.63%	95%

								networks (BNs)							

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 [13]. 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 [60,67–72]. 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's 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 [59]. 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

1. Kim D, Kim S. Sustainable Education: Analyzing the Determinants of University Student Dropout by Nonlinear Panel Data Models. Sustainability 2018;10:954. https://doi.org/10.3390/su10040954.
2. Niyogisubizo J, Liao L, Nziyumva E, Murwanashyaka E, Nshimyumukiza PC. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. Computers and Education: Artificial Intelligence 2022;3:100066. https://doi.org/10.1016/j.caeai.2022.100066.
3. Del Savio AA, Galantini K, Pachas A. Exploring the relationship between mental health-related problems and undergraduate student dropout: A

case study within a civil engineering program. Heliyon 2022;8:e09504. https://doi.org/10.1016/j.heliyon.2022.e09504.
4. Alban M, Technical University of Cotopaxi, Faculty of Computer Science and Computer Systems, Ecuador;, Mauricio D, National University of San Marcos, Artificial Intelligence Group, Perú; Predicting University Dropout trough Data Mining: A systematic Literature. Indian Journal of Science and Technology 2019;12:1-12. https://doi.org/10.17485/ijst/2019/v12i4/139729.
5. Castro R. LF, Espitia P. E, Montilla AF. Applying CRISP-DM in a KDD Process for the Analysis of Student Attrition. En: Serrano C. JE, Martínez-Santos JC, editores. Advances in Computing, vol. 885, Cham: Springer International Publishing; 2018, p. 386-401. https://doi.org/10.1007/978-3-319-98998-3_30.
6. Andrade-Girón D, Carreño-Cisneros E, Mejía-Dominguez C, Marín-Rodriguez W, Villarreal-Torres H. Comparison of Machine Learning Algorithms for Predicting Patients with Suspected COVID-19. Salud Cienc Tecnol 2023:336. https://doi.org/10.56294/saludcyt2023336.
7. Murthygowda MY, Krishnegowda RG, Venkataramu SS. Crowd Behavior Analysis and Prediction using the Feature Fusion Framework. Salud Cienc Tecnol 2022:251. https://doi.org/10.56294/saludcyt2022251.
8. Sumathi S, Gunaseelan HG. A Review of Data and Document Clustering pertaining to various Distance Measures. Salud Cienc Tecnol 2022;2:194. https://doi.org/10.56294/saludcyt2022194.
9. Tyagi S. Research Productivity on Manuscripts in the field of Social Science (2010-2020). Scopus Database. Bibliotecas Anales de Investigación 2022;18.
10. Piscitello J, Kim YK, Orooji M, Robison S. Sociodemographic risk, school engagement, and community characteristics: A mediated approach to understanding high school dropout. Children and Youth Services Review 2022;133:106347. https://doi.org/10.1016/j.childyouth.2021.106347.
11. Sletten MA, Tøge AG, Malmberg-Heimonen I. Effects of an early warning system on student absence and completion in Norwegian upper secondary schools: a cluster-randomised study. Scandinavian Journal of Educational Research 2022:1-15. https://doi.org/10.1080/00313831.2022.2116481.
12. Mikkay Ei Leen W, Jalil NA, Salleh NM, Idris I. Dropout Early Warning System (DEWS) in Malaysia's Primary and Secondary Education: A Conceptual Paper. En: Al-Emran M, Al-Sharafi MA, Shaalan K, editores. International Conference on Information Systems and Intelligent Applications, vol. 550, Cham: Springer International Publishing; 2023, p. 427-34. https://doi.org/10.1007/978-3-031-16865-9_33.
13. Chung JY, Lee S. Dropout early warning systems for high school students using machine learning. Children and Youth Services Review 2019;96:346-53. https://doi.org/10.1016/j.childyouth.2018.11.030.
14. Aljameel SS, Khan IU, Aslam N, Aljabri M, Alsulmi ES. Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients. Scientific Programming 2021;2021:1-10. https://doi.org/10.1155/2021/5587188.
15. Del Binifro F, Maurizio G, Giuseppe L, Stefano P. Predicción de la deserción estudiantil. Inteligencia artificial en la educación. 21ª Conferencia Internacional AIED 2020, Marruecos: Springer International Publishing; 2020, p. 129-40.
16. Kelleher J, Mac Namee B, D'arcy A. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press 2020.
17. Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. Third edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly; 2023.
18. Kabathova J, Drlik M. Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques. Applied Sciences 2021;11:3130. https://doi.org/10.3390/app11073130.
19. Pajankar A, Joshi A. Hands-on machine learning with python-ımplement neural network solutions with scikit-learn and pytorch. NY: Apress 2022.
20. Orooji M, Chen J. Predicting Louisiana public high school dropout through imbalanced learning techniques. 18th IEEE International Conference On Machine Learning And Applications (ICMLA), USA: IEEE; 2019, p. 456-61.
21. Caballero-Cantu JJ, Chavez-Ramirez ED, Lopez-Almeida ME, Inciso-Mendo ES, Méndez Vergaray J. El aprendizaje autónomo en educación superior. Revisión sistemática. Salud, Ciencia y Tecnología 2023;3:391. https://doi.org/10.56294/saludcyt2023391.
22. Kishore Veparala V, Kalpana V. Big Data y diferentes enfoques de clustering subespacial: De la promoción en redes sociales al mapeo genómico. Salud, Ciencia y Tecnología 2023;3:413. https://doi.org/10.56294/saludcyt2023413.
23. Kumar D, Haque A, Mishra K, Islam F, Kumar Mishra B, Ahmad S. Exploring the Transformative Role of Artificial Intelligence and Metaverse in Education: A Comprehensive Review. Metaverse Basic and Applied Research 2023;2:55. https://doi.org/10.56294/mr202355.
24. Silva-Sánchez CA. Psychometric properties of an instrument to assess the level of knowledge about artificial intelligence in university professors.

Metaverse Basic and Applied Research 2022:14. https://doi.org/10.56294/mr202214.
25. Sánchez Meca J. Cómo realizar una revisión sistemática y un meta-análisis. Aula abierta 2010, v 38, n 2 ; p 53-64 2010.
26. Serrano S, Navarro I, González M. ¿ Cómo hacer una revisión sistemática siguiendo el protocolo PRISMA?: Usos y estrategias fundamentales para su aplicación en el ámbito educativo a través de un caso práctico. Revista de pedagogía 2022;74:51-66.
27. Schwarzer G, Carpenter JR, Rücker G. Meta-Analysis with R. Cham: Springer International Publishing; 2015. https://doi.org/10.1007/978-3-319-21416-0.
28. Alexander PA. Methodological Guidance Paper: The Art and Science of Quality Systematic Reviews. Review of Educational Research 2020;90:6-23. https://doi.org/10.3102/0034654319854352.
29. Pigott TD, Polanin JR. Methodological Guidance Paper: High-Quality Meta-Analysis in a Systematic Review. Review of Educational Research 2020;90:24-46. https://doi.org/10.3102/0034654319877153.
30. Stern C, Lizarondo L, Carrier J, Godfrey C, Rieger K, Salmond S, et al. Methodological guidance for the conduct of mixed methods systematic reviews. JBI Evidence Synthesis 2020;18:2108-18. https://doi.org/10.11124/JBISRIR-D-19-00169.
31. Kiss V, Maldonado E, Segall M. The Use of Semester Course Data for Machine Learning Prediction of College Dropout Rates. Journal of Higher Education Theory and Practice 2022;22:64-74.
32. Nagy M, Molontay R. Predicting Dropout in Higher Education Based on Secondary School Performance. 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), Las Palmas de Gran Canaria: IEEE; 2018, p. 000389-94. https://doi.org/10.1109/INES.2018.8523888.
33. Rodríguez P, Villanueva A, Dombrovskaia L, Valenzuela JP. A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile. Educ Inf Technol 2023. https://doi.org/10.1007/s10639-022-11515-5.
34. Sandoval-Palis I, Naranjo D, Vidal J, Gilar-Corbi R. Early Dropout Prediction Model: A Case Study of University Leveling Course Students. Sustainability 2020;12:9314. https://doi.org/10.3390/su12229314.
35. Tan M, Shao P. Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method. Int J Emerg Technol Learn 2015;10:11. https://doi.org/10.3991/ijet.v10i1.4189.
36. Dass S, Gary K, Cunningham J. Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model. Information 2021;12:476. https://doi.org/10.3390/info12110476.
37. Kemper L, Vorhoff G, Wigger BU. Predicting student dropout: A machine learning approach. European Journal of Higher Education 2020;10:28-47. https://doi.org/10.1080/21568235.2020.1718520.
38. Aulck L, Velagapudi N, Blumenstock J, West J. Predicting Student Dropout in Higher Education 2016. https://doi.org/10.48550/ARXIV.1606.06364.
39. Wan Yaacob WF, Mohd Sobri N, Nasir SAM, Wan Yaacob WF, Norshahidi ND, Wan Husin WZ. Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques. J Phys: Conf Ser 2020;1496:012005. https://doi.org/10.1088/1742-6596/1496/1/012005.
40. Lee S, Chung JY. The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. Applied Sciences 2019;9:3093. https://doi.org/10.3390/app9153093.
41. Kashyap A, Nayak A. Different Machine Learning Models to Predict Dropouts in MOOCs. 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore: IEEE; 2018, p. 80-5. https://doi.org/10.1109/ICACCI.2018.8554547.
42. Liang J, Li C, Zheng L. Machine learning application in MOOCs: Dropout prediction. 2016 11th International Conference on Computer Science & Education (ICCSE), Nagoya, Japan: IEEE; 2016, p. 52-7. https://doi.org/10.1109/ICCSE.2016.7581554.
43. Delen D. A comparative analysis of machine learning techniques for student retention management. Decision Support Systems 2010;49:498-506. https://doi.org/10.1016/j.dss.2010.06.003.
44. Dekker G, Pechenizkiy M, Vleeshouwers J. Predicting Students Drop Out. A Case Study. International Working Group on Educational Data Mining. Educational Data Mining 2009:41-50.
45. Rodríguez-Muñiz LJ, Bernardo AB, Esteban M, Díaz I. Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? PLoS ONE 2019;14:e0218796. https://doi.org/10.1371/journal.pone.0218796.
46. Lázaro Alvarez N, Callejas Z, Griol D. Predicting Computer Engineering students' dropout in Cuban Higher Education with pre-enrollment and early performance data. J Technol Sci Educ 2020;10:241. https://doi.org/10.3926/jotse.922.
47. Yukselturk E, Ozekes S, Turel Y. Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. European Journal of Open, Distance and e-Learning 2014;17:118-33.

48. Yadav SK, Bharadwaj B, Pal S. Mining Education Data to Predict Student's Retention: A comparative Study 2012. https://doi.org/10.48550/ARXIV.1203.2987.
49. Dewan MAA, Lin F, Wen D, Kinshuk. Predicting Dropout-Prone Students in E-Learning Education System. 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), Beijing: IEEE; 2015, p. 1735-40. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.315.
50. Bayona Arévalo Y, Bolaño García M. Scientific production on dialogical pedagogy: a bibliometric analysis. Data & Metadata 2023:7. https://doi.org/10.56294/dm20237.
51. Gonzalez-Argote D. Thematic Specialization of Institutions with Academic Programs in the Field of Data Science. Data & Metadata 2023:24. https://doi.org/10.56294/dm202324.
52. Olusegun Oyetola S, Oladokun BD, Ezinne Maxwell C, Obotu Akor S. Artificial intelligence in the library: Gauging the potential application and implications for contemporary library services in Nigeria. Data & Metadata 2023;2:36. https://doi.org/10.56294/dm202336.
53. Schunck PJ. Construir el conocimiento interdisciplinar desde experiencias critico-decoloniales en educación. Salud, Ciencia y Tecnología - Serie de Conferencias 2023;2:74. https://doi.org/10.56294/sctconf202374.
54. Vergara Danies SD, Ariza Celis DC, Perpiñan Duitama LM. Strategic guidelines for intelligent traffic control. Data & Metadata 2023;2:51. https://doi.org/10.56294/dm202351.
55. Xiao T, Zhu J, Liu T. Bagging and Boosting statistical machine translation systems. Artificial Intelligence 2013;195:496-527. https://doi.org/10.1016/j.artint.2012.11.005.
56. Charles Z, Papailiopoulos D. Stability and generalization of learning algorithms that converge to global optima. International Conference on Machine Learning, s. f., p. 745-54.
57. Ying X. An Overview of Overfitting and its Solutions. J Phys: Conf Ser 2019;1168:022022. https://doi.org/10.1088/1742-6596/1168/2/022022.
58. Ghimire B, Rogan J, Galiano VR, Panday P, Neeti N. An Evaluation of Bagging, Boosting, and Random Forests for Land-Cover Classification in Cape Cod, Massachusetts, USA. GIScience & Remote Sensing 2012;49:623-43. https://doi.org/10.2747/1548-1603.49.5.623.
59. Yaman E, Subasi A. Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Automated EMG Signal Classification. BioMed Research International 2019;2019:1-13. https://doi.org/10.1155/2019/9152506.
60. Altman N, Krzywinski M. Ensemble methods: bagging and random forests. Nature Methods 2017:933-5.
61. Bacigalupe MDLA. Emociones y movimiento en el estudio inter(trans)disciplinario del comportamiento humano desde dentro. Salud, Ciencia y Tecnología - Serie de Conferencias 2023;2:83. https://doi.org/10.56294/sctconf202383.
62. Gamboa Rosales NK, Celaya-Padilla JM, Galván-Tejada CE, Galván-Tejada JI, Luna-García H, Gamboa-Rosales H, et al. Infotainment technology based on artificial intelligence: Current research trends and future directions. Iberoamerican Journal of Science Measurement and Communication 2022;2. https://doi.org/10.47909/ijsmc.144.
63. Jiménez-Pitre I, Molina-Bolívar G, Gámez Pitre R. Systemic vision of the technological educational context in Latin America. Region Cientifica 2023:202358. https://doi.org/10.58763/rc202358.
64. Laplagne Sarmiento C, Urnicia JJ. B-learning protocols for information literacy in Higher Education. Region Cientifica 2023:202373. https://doi.org/10.58763/rc202373.
65. Silva Júnior EMD, Dutra ML. A roadmap toward the automatic composition of systematic literature reviews. Iberoamerican Journal of Science Measurement and Communication 2021;1:1-22. https://doi.org/10.47909/ijsmc.52.
66. Kavzoglu T, Teke A. Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost). Arab J Sci Eng 2022;47:7367-85. https://doi.org/10.1007/s13369-022-06560-8.
67. Basantes E, Ortega C, Valle V. Innovadora gestión del conocimiento para el aprendizaje cooperativo en la Educación Básica Superior. Bibliotecas Anales de Investigación 2023;19.
68. Musiño C, Alvarado J. Las metodologías aplicadas en los artículos científicos de las Ciencias Bibliotecaria y de la Información, y Big Data. Bibliotecas Anales de Investigación 2021;17.
69. Tiwari P, Chaudhary S, Majhi D, Mukherjee B. Comparing research trends through author-provided keywords with machine extracted terms: A ML algorithm approach using publications data on neurological disorders. Iberoamerican Journal of Science Measurement and Communication 2023;3. https://doi.org/10.47909/ijsmc.36.
70. Takaki P, Dutra M. Data science in education: interdisciplinary contributions. En: Rodrigues Dias TM, editor. Advanced Notes in Information Science, vol. 2, ColNes Publishing; 2022. https://doi.org/10.47909/anis.978-9916-9760-3-6.94.
71. Ruiz-Mori I, Romero-Carazas R, Espíritu-Martínez A, Mamani-Jilaja D, Valero-Ancco N, Flores-Chambilla S. Análisis bibliométrico de la

producción científica sobre competencia y brecha digitales. Bibliotecas Anales de Investigación 2023.
72. Zaina R, Ramos VFC, De Araujo GM. Automated triage of financial intelligence reports. En: Rodrigues Dias TM, editor. Advanced Notes in Information Science, vol. 2, ColNes Publishing; 2022. https://doi.org/10.47909/anis.978-9916-9760-3-6.115.
73. Adetunji AB, Akande ON, Ajala FA, Oyewo O, Akande YF, Oluwadara G. House Price Prediction using Random Forest Machine Learning Technique. Procedia Computer Science 2022;199:806-13. https://doi.org/10.1016/j.procs.2022.01.100.
74. Albarracín Vanoy RJ. STEM Education as a Teaching Method for the Development of XXI Century Competencies. Metaverse Basic and Applied Research 2022:21. https://doi.org/10.56294/mr202221.
75. Catrambone R, Ledwith A. Enfoque interdisciplinario en el acompañamiento de las trayectorias académicas: formación docente y psicopedagógica en acción. Interdisciplinary Rehabilitation / Rehabilitación Interdisciplinaria 2021;3.
76. Junco Luna G. Study on the impact of artificial intelligence tools in the development of university classes at the school of communication of the Universidad Nacional José Faustino Sánchez Carrión. Metaverse Basic and Applied Research 2023;2:51. https://doi.org/10.56294/mr202351.
77. Nahi HA, Asaad Hasan M, Hussein Lazem A, Ayad Alkhafaji M. Securing Virtual Architecture of Smartphones based on Network Function Virtualization. Metaverse Basic and Applied Research 2023:37. https://doi.org/10.56294/mr202337.
78. Simhan L, Basupi G. None Deep Learning Based Analysis of Student Aptitude for Programming at College Freshman Level. Data & Metadata 2023;2:38. https://doi.org/10.56294/dm202338.
79. Malek N, Yaacob W, Wah Y, Md Nasir S, Shaadam N, Indratno S. Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data. IJEECS s. f.;29.
80. Pu L, Shamir R. 4CAC: 4-class classifier of metagenome contigs using machine learning and assembly graphs. Bioinformatics; 2023. https://doi.org/10.1101/2023.01.20.524935.