Journal of Computational Social Science (2024) 7:171–196
https://doi.org/10.1007/s42001-023-00231-w

SURVEY ARTICLE

Exploring statistical approaches for predicting student dropout in education: a systematic review and meta-analysis

Raghul Gandhi Venkatesan[1] · Dhivya Karmegam[2] · Bagavandas Mappillairaju[3]

Received: 17 June 2023 / Accepted: 23 October 2023 / Published online: 29 November 2023
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd. 2023

Abstract

Student dropout is non-attendance from school or college for an extended period for no apparent cause. Tending to this issue necessitates a careful comprehension of the basic issues as well as an appropriate intervention strategy. Statistical approaches have acquired much importance in recent years in resolving the issue of student dropout. This is due to the fact that statistical techniques can efficiently and effectively identify children at risk and plan interventions at the right time. Thirty-six studies in total were reviewed to compile, arrange, and combine current information about statistical techniques applied to predict student dropout from various academic databases between 2000 and 2023. Our findings revealed that the Random Forest in 23 studies and the Decision Tree in 16 studies were among the most widely adopted statistical techniques. Accuracy and Area Under the Curve were the frequently used evaluation metrics that are available in existing studies. However, it is notable that the majority of these techniques have been developed and tested within the context of developed nations, raising questions about their applicability in different global settings. Moreover, our meta-analysis estimated a pooled proportion of overall dropouts of 0.2061 (95% confidence interval: 0.1845–0.2278), revealing significant heterogeneity among the selected studies. As a result, this systematic review and meta-analysis provide a brief overview of statistical techniques focusing on strategies for predicting student dropout. In addition, this review highlights unsolved problems like data imbalance, interpretability, and geographic disparities that might lead to new research in the future.

Keywords Dropout prediction · Supervised learning · Unsupervised learning · Meta-analysis

Extended author information available on the last page of the article

Introduction

There is an ever-increasing interest in exploring the subject of education dropout worldwide, with high incidence risks as one of the biggest challenges [1]. Dropout seriously impacts education systems, resulting in lower enrolment and failure to meet academic goals [2]. As a result, schools, colleges, universities, and governments face economic and social consequences. Moreover, when administrators lack the resources to detect at-risk students in danger of dropping out, dropout becomes a severe subject [3]. Consequently, only some remedial procedures are adapted on time to retain students in schools and colleges [4]. So, predicting student dropout and detecting the elements that could lead to this significant phenomenon is now becoming a priority [5]. Most of the predictive modeling techniques used need to be explained. This might be one of the reasons why this issue of dropouts still exists [1].

Machine Learning (ML) techniques are among the most widely researched solutions for dropout prediction. In developed nations, extensive research has been done on creating student dropout prediction algorithms [6–8]. Furthermore, there is substantial work on ML-based techniques to prevent dropouts [9, 10]. Literature-based knowledge can shift the dropout prevention effort from responsive to proactive. This could be more practical now than at any other time because Information and Communication Technologies (ICTs) have effectively changed how information has been gathered and handled, a vital element to the data-driven harnessing of a logical sequence of observed occurrences. However, confusion still prevails concerning the viability of the current insightful procedures and models. Despite a few past research endeavors, difficulties still need to be addressed.

The necessity to check the efficiency of systematic reviews examining the bold prediction of education dropout using statistical techniques has inspired us to investigate. In addition, this review looks beyond the results to highlight persistent problems such as data imbalance, interpretability, and geographic inequalities. These unresolved problems are noted as promising areas for future study, highlighting the domain’s dynamism and the chances for advancement and innovation in dealing with student dropout on a worldwide scale. A sequential procedure is used to recognize, select, and evaluate the synthesized investigation results to align with research objectives [11, 12]. This systematic review and meta-analysis aim to survey the statistical techniques-based works conducted in education dropout prediction between 2000 and 2023. The objectives of the study are:

1. To better understand the statistical methods and strategies used to predict student dropout.
2. To evaluate the efficiency and standard performance metrics of current statistical methods in the reviewed studies.
3. To recognize the exploration difficulties and limits confronting the current statistical techniques for predicting student dropout.

Methods

Survey methodology

This systematic review and meta-analysis were performed to examine what types of statistical techniques are used to predict early education dropout. We framed the Population, Intervention, Comparison, Outcome (PICO) model to justify the above research question. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for searching and assessing published articles for systematic review and meta-analysis are followed [13].

Search strategy

The electronic literature databases IEEE Explore, ScienceDirect, Web of Science, Association for Computing Machinery (ACM), and Scopus were searched to collect the relevant articles published between 2000 and 2023. Figure 1 summarizes the general procedures followed in our systematic review. The databases were searched using the identified keywords. The keywords are trailed multiple times and modified in each database to obtain relevant studies and are given in Appendix A. Finally, full-text papers were manually searched and selected for this study.

Inclusion and exclusion criteria

Research articles that are obtained in the search strategy are assessed using the inclusion criteria fixed by researchers. The research articles published in peer-reviewed journals or scientific forums were mainly considered. Studies that predict early education dropout using statistical techniques were considered for review. At the same time, a study that uses student databases for the prediction was also included. Studies that do not focus on the student's education dropout prediction were excluded from the review. Studies missing the requisite data for a thorough investigation of dropout prediction were also eliminated. As part of this systematic review, we followed PRISMA's basic principles for ensuring clarity and quality. Table 1 provides the criteria for inclusion that we used when selecting the articles.

Two authors independently reviewed the potential research studies by reading the titles and abstracts of the articles using the search query and manual searches. In the second phase, the identified research articles are completely screened to remove duplicates and irrelevant studies. Two authors examine the selected research papers independently to decide whether to include the research paper in the review. A third author solved discrepancies between the two authors through a joint discussion. The included articles' quality is evaluated using the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Analytical Cross-Sectional studies in the third phase [14]. JBI is a methodological quality assessment tool for various types of research. Based on this assessment, methodologically good articles were included for qualitative synthesis. The JBI checklist of studies in this review did not exclude many

Fig. 1 Steps of our review methodology

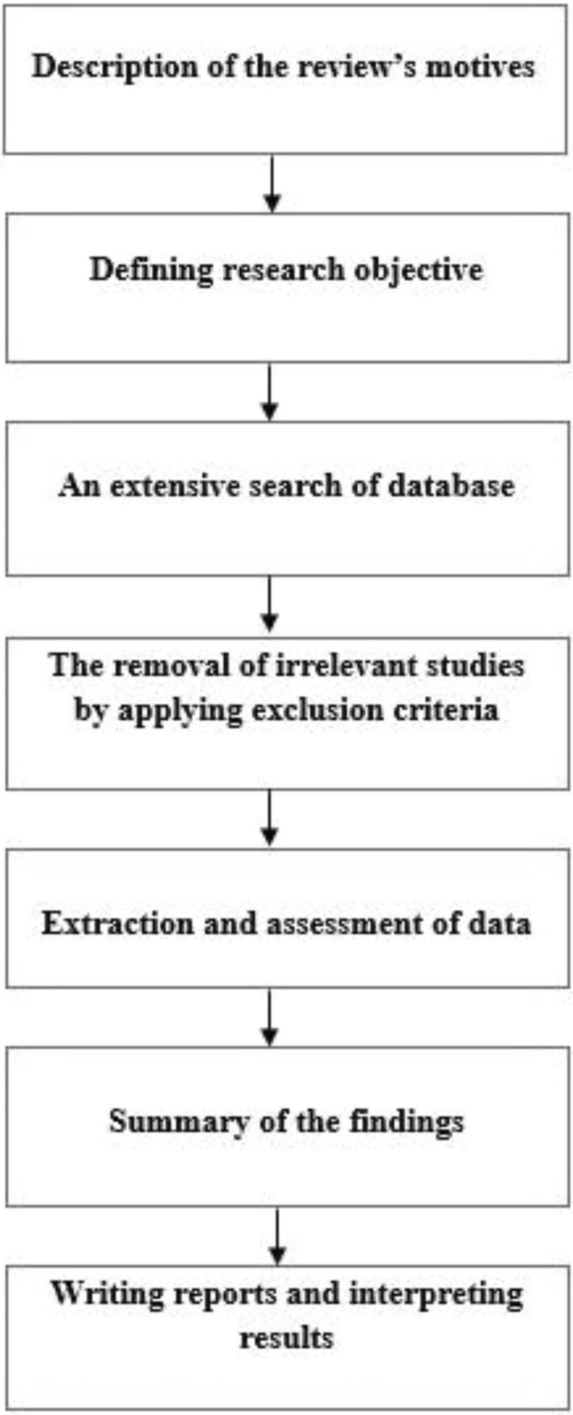

Description of the review's motives
Defining research objective
An extensive search of database
The removal of irrelevant studies by applying exclusion criteria
Extraction and assessment of data
Summary of the findings
Writing reports and interpreting results

Table 1 Inclusion criteria in our systematic literature review

Inclusion criteria	Description
Aim	Studies that predict early education dropout using statistical techniques
Empirical evidence of prediction	Studies with dropout databases should be used for early dropout prediction
Language of publication	English written research studies
Year of publication	Research studies published from 2000 to 2023
Publication details	Studies must be published in scientific forums
Availability of articles	Open-access and full-text articles will be considered for qualitative synthesis

owing to poor methodology. A summary of the risk of bias is provided in Appendix B. As shown in Fig. 2, we have searched and screened articles using the PRISMA flowchart.

Data extraction

An effective data extraction design was created in Microsoft Excel based on the survey objectives and the inclusion criteria [15]. Research articles selected for qualitative synthesis were analyzed to extract the data supporting the review's primary focus. Information such as types of dropouts, country, sample size, data source and software used, methodology, the prevalence of the study, and year of publication and title of the study were extracted.

The initial records identified through a database search of IEEE Explore included 749 articles, Science Direct 51 articles, Association for Computing Machinery (ACM) 453 articles, Scopus 185 articles, and Web of Science 858 articles. After screening the title and abstract of each research article, 29 articles from IEEE Explore, 6 articles from Science Direct, 10 articles from ACM, 23 articles from Scopus, and 25 articles from Web of Science were selected for further investigation. A Manual search resulted in the inclusion of 20 articles. Removing duplicates and screening full texts included 13 articles from IEEE Explore, 5 articles from Science Direct, 3 articles from ACM, 2 articles from Scopus, 6 articles from Web of Science, and 7 articles through manual search. Appendix C summarizes the outcomes of the search and screening details. The data for both qualitative and quantitative synthesis were taken from 36 and 22 publications finalized. Table 2 provides the characteristics of the included studies, and other features are given in Appendix D.

Statistical analysis

To determine the proportion of the minority class (dropouts), a PRAW (meta-analysis with random effects and a summary measure of proportion random effects model) analysis was conducted. Multiple factors led to the selection of the PRAW method. First, it is incredibly well suited for meta-analytic examination of

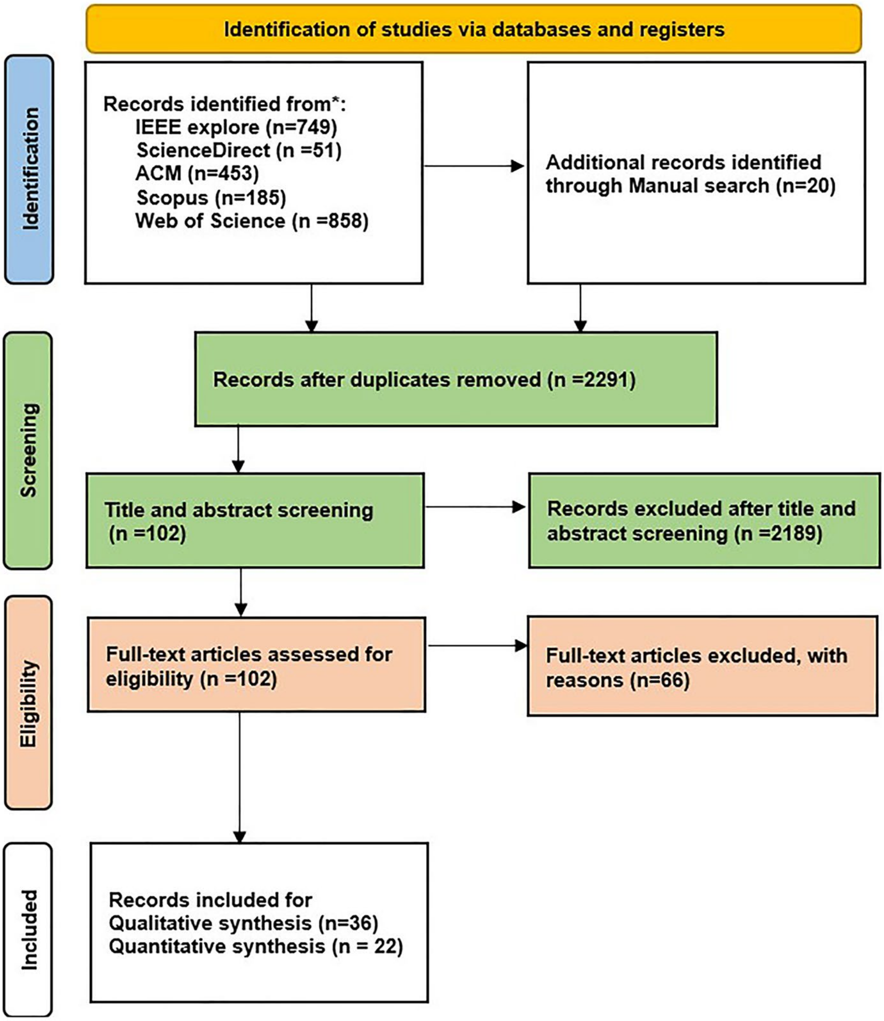

Identification of studies via databases and registers
Identification
Records identified from*:
IEEE explore (n=749)
ScienceDirect (n =51)
ACM (n=453)
Scopus (n=185)
Web of Science (n =858)
Additional records identified through Manual search (n=20)
Screening
Records after duplicates removed (n =2291)
Title and abstract screening (n =102)
Records excluded after title and abstract screening (n =2189)
Eligibility
Full-text articles assessed for eligibility (n =102)
Full-text articles excluded, with reasons (n=66)
Included
Records included for
Qualitative synthesis (n=36)
Quantitative synthesis (n = 22)

Fig. 2 PRISMA flowchart of our systematic review

binary outcomes, like dropouts vs. non-dropouts. Second, PRAW is robust and suitable to our analysis because it considers both within-study and between-study variability, where several studies from various sources were included [51]. Furthermore, using random effects models enables a more accurate representation of the underlying diversity among multiple studies. The I^2 statistic, with an I^2 value between 75 and 100%, reflects the variability among research. To determine the reliability of our conclusions, we performed a sensitivity analyses by established practices for meta-analysis using leave-one-out approach. We further stratified our analysis by type of dropout (University vs. School). Additionally, publication bias was carefully examined using trim and fill plots, Egger's test, and rank

Table 2 Features of the included studies

S. no	Year	First author	Name of dropout	Country	Sample size	Data source	Data collection period	Software
1	2012	Carl Lamote [16]	School	Netherland	4735	Longitudinal research in secondary education (LOSO)	NA	NA
2	2013	Carlos Marquez Vera [17]	School	North America	670	Unit preparatory at the autonomous university of Zacatecas	2009–2010	Weka
3	2015	Ara NB [18]	College	Denmark	72,598	MaCom Lectio database	2009	Weka
4	2016	S.Ameri [10]	University	North America	11,121	Wayne state university	2002–2009	R and Weka
5	2017	Evandro B. Costa [19]	University	Brazil	262 and 161	Brazilian Public University	2013 & 2014	Pentaho Data Integration Tool & Weka
6	2017	Lovenoor Aulck [20]	University	North America	32,538	University of Washington	2013	NA
7	2017	Elizabeth H. Weybright [21]	School	Africa	601	NA	2004–2007	Statistical Analysis System (SAS)
8	2018	Melissa Adelman [22]	School	North America	19,000	Guatemala and Honduras Administrative Data	2013	NA
9	2018	Johannes Berens [23]	University	Germany	14,496 and 7600	federal state of North Rhine-Westphalia & private university of applied sciences	2007–2017	NA
10	2018	Vinayak Hegde [24]	University	India	50	Amrita school of arts and sciences, Mysuru	NA	R and Weka
11	2018	Marcell Nagy [25]	University	Hungary	15,825	Budapest University of Technology and Economics	2010–2017	Rapid Miner

Table 2 (continued)

S. no	Year	First author	Name of dropout	Country	Sample size	Data source	Data collection period	Software
12	2019	Sunbok Lee [26]	High School	South Korea	1,65,715	National Education Information System (NEIS), South Korea	2014	R
13	2019	Paulo.M.da Silva [27]	University	Brazil	1,33,528	National Institute of Educational Studies and Research	2013	NA
14	2019	Raghad Al-Shabandar [28]	University	North America	11,300	Harvard University, MIT	2013–2014	NA
15	2019	Thiago M. Barros[29]	School	Brazil	5788	Unified Public Administration System	2018	Python
16	2019	Al Amin Biswas [30]	University	Bangladesh	66	Public University of Bangladesh	2012–2017	NA
17	2020	Francesca Del Bonifro [31]	University	Italy	15,000	11 schools of university	2016–2017	Python
18	2020	Warit Tenpipat [32]	University	Thailand	13,714	King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, Thailand	2012—2020	NA
19	2020	Frederick F. Patacsil [33]	University	Phillippines	2,401	Pangasinan State University—Urdaneta City Campus	2012–2016	Rapid Miner
20	2020	Máté Baranyi [34]	University	Hungary	8,319	Budapest University of Technology and Economics	2013–2019	Python
21	2020	Sagarika Nangia [35]	School	India	17,359	Unified District Information System for Education	2016–2017	NA

Table 2 (continued)

S. no	Year	First author	Name of dropout	Country	Sample size	Data source	Data collection period	Software
22	2020	Roderick Lottering [36]	University	South Africa	4419	Tshwane University of Technology	2013–2017	NA
23	2020	Francisco A. da S. Freitas [37]	University	Brazil	1549	Federal Institute of Education, Science, and Technology of Ceara (IFCE)	2008–2019	Python, Java, SQL
24	2021	Sebastian ´ Maldonado [38]	University	South America	3,290	Chilean University	2012–2016	R
25	2021	Diego Opazo [39]	University	South America	31,714 and 73,067	Chilean University	NA	Python
26	2021	Antonio Jesús Fernández-García [40]	University	Spain	1418	Spanish University	2012–2019	Python
27	2022	Emanuel Marques Queiroga [41]	School	Uruguay	261,446	National Administration of Public Education (ANEP)	2015–2020	Python
28	2022	Marina Segura [42]	University	Spain	3428	Spanish Public University	2017–2018	NA
29	2022	Diogo E. Moreira da Silva [43]	University	Portugal	331	Universidade de Trás-os-Montes e Alto Douro (UTAD)	2011–2019	NA
30	2022	Yuda N Mnyawami [44]	School	Africa	385,634	Twaweza Uwezo data repository	NA	Python
31	2022	Delali Kwasi Dake [45]	University	Ghana	1239	University of Education, Winneba	NA	Weka
32	2022	Jovial Niyogisubizo [46]	University	Slovakia	261	Constantine the Philosopher University in Nitra	2016–2020	NA
33	2022	Vaneza Flores [47]	University	Peru	4365	National University of Moquegua (UNAM)	2008–2019	Weka

Table 2 (continued)

S. no	Year	First author	Name of dropout	Country	Sample size	Data source	Data collection period	Software
34	2023	Daniel A. Gutierrez-Pachas [48]	University	South America	13,696	Latin American University	2008–2020	Python
35	2023	kamal Samy Selim [49]	School	Egypt	10,916	Survey of Young people in Egypt (SYPE)	2014	Python
36	2023	Zihan Song [50]	University	Korea	60,010	Korean University	2010–2021	NA

**NA* not applicable

correlation test. The meta and metafor packages were used for all analyses in R Studio.

Result

This section contains basic information on the reviewed studies, the data pre-processing techniques used, the statistical methods used to predict student dropout, and the model assessment metrics used to evaluate the performance of the model.

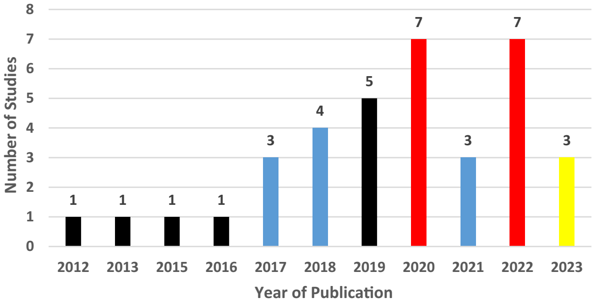
Trends in dropout prediction research

The quantity of studies focussing on predicting student education dropouts is steadily increasing. Since 2017, the interest in dropout prediction models has increased, which depicts the universal trend in identifying at-risk students early to improve their proficiency in education, as shown in Fig. 3. Our database search turned up published studies to the end of March 2023, which might indicate a minor drop in the count of published studies in 2023.

Geographic distribution of research

The result compiled research that was carried out in different nations such as Bangladesh, Brazil, Chile, Denmark, Egypt, Germany, Ghana, Guatemala, Hungary, India, Italy, Korea, the Netherlands, Peru, Philippines, Portugal, Slovakia, South Africa, South Korea, Spain, Thailand, United States of America, and Uruguay. We have opted to group them into continents, as shown in Fig. 4. Ten studies were from Europe, followed by nine studies in South America, Seven in Asia, and five in Africa and North America. Out of 36 studies, twenty-six studies (72%) used University databases to predict student dropout. Ten studies (28%) used school dropout data, as shown in Fig. 5.

Fig. 3 Number of articles distributed by year of publication



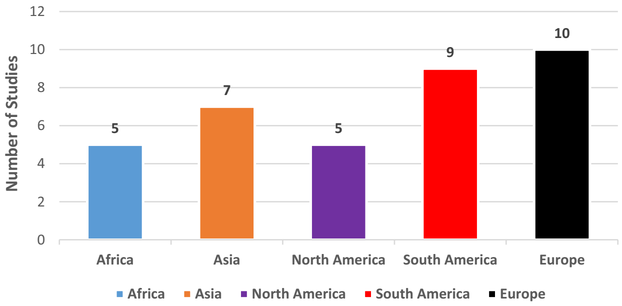
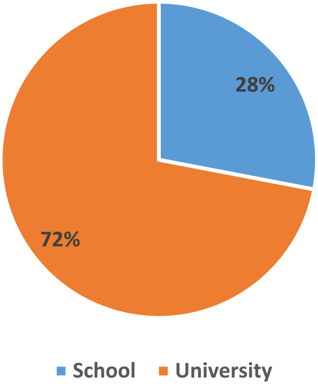


Fig. 4 Distribution of included studies by continents

Fig. 5 Distribution of included studies by type of dropout



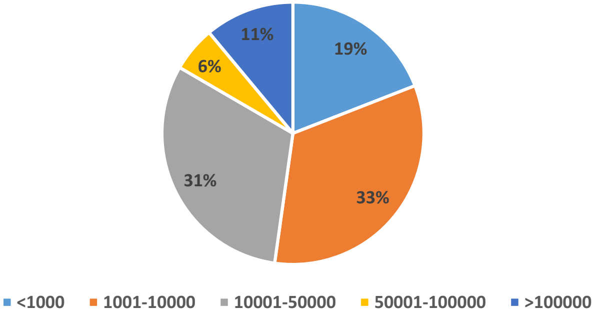


Fig. 6 Distribution of included studies by sample size

Dataset characteristics

The datasets used in the research articles are classified into five categories based on the number of data included, as mentioned in Fig. 6. Seven studies used an experimental dataset of less than 1000 students, twelve used datasets ranging from 1001 to 10,000 students, and eleven used 10,001–50,000 students. The remaining six studies used more than 50,000 student datasets for early prediction. When we analyzed the prediction accuracy of each prediction model, we found varied results based on the sample size used in their study. The articles with smaller datasets as well as larger datasets have provided mixed findings. To iterate, the included study with a sample size of 2401 students gave a weak prediction as the accuracy of the model was found to be (70.47%)[33]. In contrast, a study conducted in Slovakia with 261 students resulted in a prediction with greater accuracy of 91.66% [46].

In this context, it is essential to point out that arriving at a significant conclusion based on the classification of training datasets into two pools, small or adequate sample size, is a challenging task. To reason, the results derived from mixed findings are influenced by factors such as data imbalance, error tolerance, and the kind of prediction technique used. Also, comparing the performance of each prediction model on varying datasets will not provide conclusive results. There is no conflict that the greater the sample, the more accurate the prediction. Whatever the case, this was not obvious from our qualitative synthesis.

Software utilization in dropout prediction

As for the software used to examine the datasets, we recognize eight different software from 22 studies in Fig. 7. The outcome features of broadly utilized

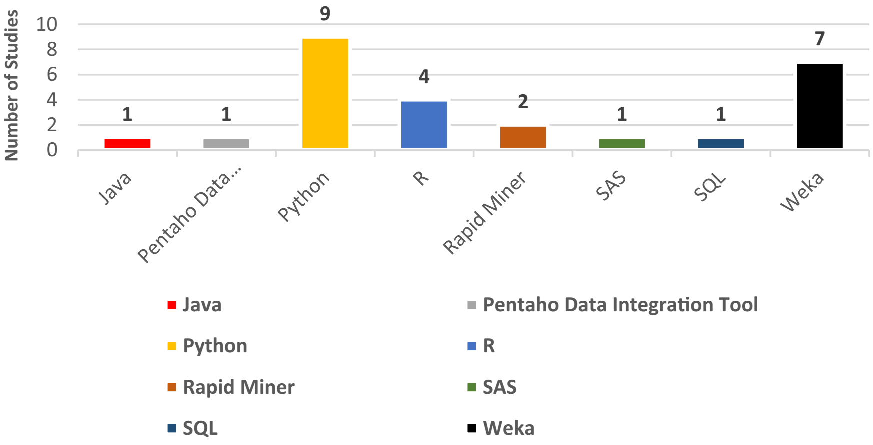


Fig. 7 Distribution of included studies by software used for analysis

software WEKA, R, and Python are doubtless because of their wide range of programmed learning algorithms for mining, adaptability in modeling, and functionalities. Fifteen investigations should have indicated the software they used to develop their dropout models.

Understanding prediction methods

Data pre-processing

The essential target of information pre-handling is comprehending the information and its factors. Before applying statistical techniques, it is vital to do some pre-processing procedures such as cleaning, integration, encoding, attribution, dimensionality decrease, standardization, and variable transformations. The quality and dependability of accessible data directly influence the outcome acquired; hence, pre-processing should be highlighted as a significant task. In practice, certain specific pre-processing procedures were used to set up every one of the recently portrayed data to complete grouping assignments accurately. First, all accessible data were incorporated into a solitary dataset. Those students who did not have 100 percent full data were wiped out throughout this cycle.

As one can expect, the dataset is profoundly unequal since the students who leave the studies are a minority, and the proportion between the negative (non-dropout) and positive (dropout) models is around 3:1. Even though this is great for educational management, preparing an ML model for paired order with an exceptionally unbalanced dataset may result in poor final performance, fundamentally in light of the fact that in such a situation the classifier would underrate the class with a lower number of tests [52]. To resolve this issue, sampling or balancing/rebalancing algorithms may be applied to the data prior to pre-processing.

Only a few studies in this qualitative synthesis have used the Synthetic Minority Oversampling Technique (SMOTE), Random under-sampling, and Random over-sampling. Appendix E shows 26 different data pre-processing techniques found in this survey.

Data analysis

After pre-processing the acquired data, extraction, and analysis were used to transform the data into information and achieve the desired outputs. Regarding education, statistical or machine-learning approaches can be used to estimate student dropout.

Supervised or structured learning depends on training from a collection of labeled data in the test dataset as it can distinguish unlabelled data in the test set to the maximum possible accuracy [53]. The worldview of this learning is effective, and it generally finds answers for a few linear and nonlinear problems, such as classifications, predictions, forecasts, advanced mechanics, and so on.

Previous research concentrated on supervised or structured learning techniques for identifying academic dropout students. For example, the commonly used models are Bayesian Classifier, Association Rule Learning, Logistic Regression, Random

Forest, Ensemble Learning, and Neural Network models [54]. As for clustering approaches, the researchers prefer neural networks and decision trees for forecasting students' success [55]. Gray et al. 2014 describe a neural network as having the unique feature of recognizing all the possible correlations between indicator factors and being able to detect independent and dependent factors with no doubt [56, 57]. In contrast, decision trees have been employed to discover smaller or larger data structures and forecast their value as they are simple and straightforward [58, 59].

A few predictive models were developed to resolve the issue of dropout utilizing various methodologies like time-to-event analysis, the Generalized Linear Model, Linear Discriminant Analysis, and Bayesian Network [10, 31, 47]. Different Regression methodologies, such as Probit Regression, Multi-Task Logistic Regression, and Neural Multi-Task Logistic Regression, were also introduced to perform early student dropout identification. The utilization of unsupervised learning is not found in any of our chosen investigations.

The time-to-event analysis is utilized to investigate information from the time until the incident occurs [60]. It provides different components to deal with censored data issues that emerge in modeling the longitudinal data, which happens universally in other application spaces [48].

The use of time-to-event analysis to examine student dropout was pioneered in the area of education and management. To investigate the impact of various school classifications on the school effects, Carl Lamote et al. (2013) employed a multilevel discrete-time hazard model [16]. Ameri et al. (2016) used a semi-parametric method to construct a time-to-event analysis framework to determine in-danger pupils [10]. This methodology collects time-varying characteristics and exploits that knowledge to estimate student dropout better, utilizing the available students datasets from 2002 to 2009 of Wayne State University. Indeed, in time-to-event analysis, individuals are generally monitored for a specific amount of time, focusing on the moment the event of interest happens [61]. Analyzing academic, socioeconomic, and equity factors, Daniel A. Gutierrez-Pachas et al. (2023) employed parametric, semi-parametric, and advanced survival approaches to predict higher education dropout [48]. As a result, the advantage of time-to-event analysis over the other techniques is the potential to incorporate a temporal factor into the framework as well as to manage censored information successfully. Despite the fact that the effectiveness of time-to-event analysis approaches in other fields, such as health sectors, technology, finance, human resource management, etc., there need to be more attempts to apply such techniques to the problem of student dropout [62].

Linear Discriminant Analysis acts as a dimensional reduction algorithm attempting to lessen the data complexity by projecting the real component space on a lower-dimensional one while attempting to hold significant data variation; likewise, it doesn't include parameter settings. Del Bonifro et al. (2020) fostered an ML technique to anticipate the dropout of a first-year undergraduate student. The proposed technique permits estimating the danger of leaving a scholarly course, and it tends to be utilized either during the application stage or during the principal year [31].

The Exemplary methodology for estimating a statistical connection between an independent variable and a few other independent variables is the regression technique. Berens et al. fostered an early recognition framework utilizing probit

Table 3 Frequency of most involved prediction techniques

Model name	Frequency
Random Forest	23
Decision Tree	16
Logistic Regression	14
Support Vector Machine	12
Artificial Neural Network	11
Naïve Bayes	10
K-Nearest Neighbour	10

regression to predict student success in tertiary education and provide designated intervention [23]. Appendix F categorizes the 36 research articles as per the prediction methods utilized for early student dropout. Table 3 shows the most often involved prediction strategies in the qualitative synthesis. Random Forest has been used in 23 studies, followed by a Decision tree in 16 studies, Logistic Regression in 14 studies, Support Vector Machine in 12 studies, Artificial Neural Network in 11 studies, Naïve Bayes classifier, and K-Nearest Neighbour in 10 studies.

Performance evaluation

Evaluation metrics for dropout prediction

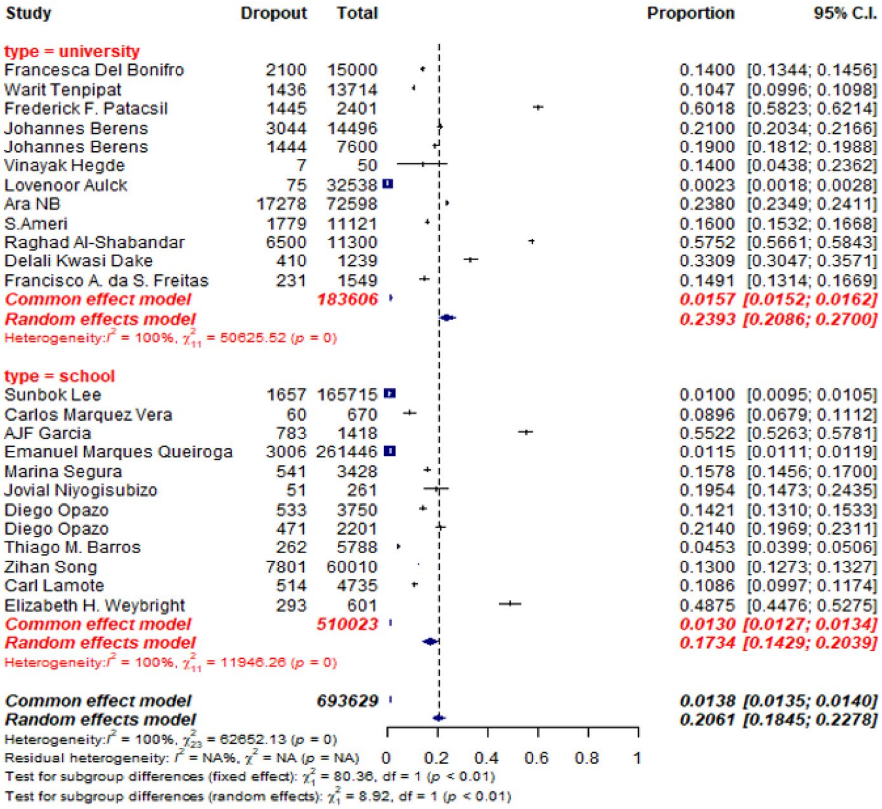
Given the inherent probabilistic nature of predictive models, evaluating the outcomes while using them is crucial. A model's performance is greatly influenced by evaluation measures, which also help determine what changes should be made to improve the accuracy of predictions. Various criteria have been suggested in the literature to achieve more accurate prediction models [63, 64]. The kind of problem being handled determines the evaluation metrics to be used. We analyzed the model measurements used in the investigations and found that 21 investigations evaluated their prediction quality through 'accuracy,' followed by area under the curve (AUC) in 19 studies. Precision and Recall, kappa, and F1 measures are additional performance measures identified for classification problems [50].

Similarly, Mean Squared Error (MSE) and Mean Absolute Error (MAE) were identified in fewer studies for regression problems [10, 27, 48]. The reviewed studies used several performance indicators to assess the validity of the dropout prediction models. A single measure was utilized in eight (22%) investigations to evaluate the dropout prediction. Seven investigations (19%) employed two performance criteria, whereas ten investigations (28%) used three. More than four measures were used in 37% of the investigations given in Appendix D.

Challenges in predicting student dropout

Data imbalance

We meta-analyzed the 22 studies that reported the proportion of dropouts in a dataset out of the 36 papers that were included in the qualitative synthesis. The estimated pooled proportion of overall dropouts was found to be 0.2061 (95% confidence interval (CI): 0.1845–0.2278). Estimates varied greatly from 0.0023 to 0.6018, which might be partially attributed to variations in dropout types. This means that, on average, only 20% of dropout samples were used for dropout prediction. This indicates the data imbalance problem in the prediction of student dropout problems. The estimated tau-squared value was found to be 0.0016, with a standard error (SE) of 0.0015, suggesting some heterogeneity among the effect sizes of included studies. The I^2 statistic also showed significant heterogeneity with an I^2 value of 99.96% ($P<0.05$). We carried out further stratified meta-analyses to better comprehend this component's interaction. Focusing on the specific type of dropout, we calculated a pooled proportion of dropouts in universities to be 0.2393 (95% CI: 0.2086–0.2700), which was lower than



Study	Dropout	Total	Proportion	95% C.I.
type = university				
Francesca Del Bonifro	2100	15000	0.1400	[0.1344; 0.1456]
Warit Tenpipat	1436	13714	0.1047	[0.0996; 0.1098]
Frederick F. Patacsil	1445	2401	0.6018	[0.5823; 0.6214]
Johannes Berens	3044	14496	0.2100	[0.2034; 0.2166]
Johannes Berens	1444	7600	0.1900	[0.1812; 0.1988]
Vinayak Hegde	7	50	0.1400	[0.0438; 0.2362]
Lovenoor Aulck	75	32538	0.0023	[0.0018; 0.0028]
Ara NB	17278	72598	0.2380	[0.2349; 0.2411]
S.Ameri	1779	11121	0.1600	[0.1532; 0.1668]
Raghad Al-Shabandar	6500	11300	0.5752	[0.5661; 0.5843]
Delali Kwasi Dake	410	1239	0.3309	[0.3047; 0.3571]
Francisco A. da S. Freitas	231	1549	0.1491	[0.1314; 0.1669]
Common effect model		*183606*	*0.0157*	*[0.0152; 0.0162]*
Random effects model			*0.2393*	*[0.2086; 0.2700]*
Heterogeneity: I^2 = 100%, χ^2_{11} = 50625.52 (p = 0)				
type = school				
Sunbok Lee	1657	165715	0.0100	[0.0095; 0.0105]
Carlos Marquez Vera	60	670	0.0896	[0.0679; 0.1112]
AJF Garcia	783	1418	0.5522	[0.5263; 0.5781]
Emanuel Marques Queiroga	3006	261446	0.0115	[0.0111; 0.0119]
Marina Segura	541	3428	0.1578	[0.1456; 0.1700]
Jovial Niyogisubizo	51	261	0.1954	[0.1473; 0.2435]
Diego Opazo	533	3750	0.1421	[0.1310; 0.1533]
Diego Opazo	471	2201	0.2140	[0.1969; 0.2311]
Thiago M. Barros	262	5788	0.0453	[0.0399; 0.0506]
Zihan Song	7801	60010	0.1300	[0.1273; 0.1327]
Carl Lamote	514	4735	0.1086	[0.0997; 0.1174]
Elizabeth H. Weybright	293	601	0.4875	[0.4476; 0.5275]
Common effect model		*510023*	*0.0130*	*[0.0127; 0.0134]*
Random effects model			*0.1734*	*[0.1429; 0.2039]*
Heterogeneity: I^2 = 100%, χ^2_{11} = 11946.26 (p = 0)				
Common effect model		*693629*	*0.0138*	*[0.0135; 0.0140]*
Random effects model			*0.2061*	*[0.1845; 0.2278]*

Heterogeneity: I^2 = 100%, χ^2_{23} = 62652.13 (p = 0)

Residual heterogeneity: I^2 = NA%, χ^2 = NA (p = NA)

Test for subgroup differences (fixed effect): χ^2_1 = 80.36, df = 1 (p < 0.01)

Test for subgroup differences (random effects): χ^2_1 = 8.92, df = 1 (p < 0.01)

Fig. 8 Forest plot showing the pooled proportion of dropout

school dropouts 0.1734 (95% CI: 0.1429–0.2039), which showed a significant difference. We observe that estimates for these groups ranged widely, from 0.0023 to 0.6018 for University dropouts to 0.0100 to 0.5522 for school dropouts, as shown in Fig. 8. Sensitivity analysis is performed to strengthen the reliability of the findings using the Baujot plot, Influencer analysis, and Leave-one-out analysis. The results of the trim and fill method, rank correlation test for funnel plot asymmetry showed no publication bias, and the Eggers test revealed a publication bias, which is given in Appendix G.

Discussion

Student dropout prediction has become essential for higher education pioneers [65]. Statistical learning has acquired massive momentum in the past ten years to improve early student dropout identification. Statistical learning and advanced machine learning are declared to work on fulfilling dropout prediction [66]. Various significant concerns exist for computerizing the appraisal of individual dropouts that might serve as a bridge to student success in school [67, 68]. Yet, it is still being determined how statistical learning and machine learning can be used to illustrate and predict early student dropout. The present systematic review was conducted to make an effort to connect this gap in research.

To answer the first and second objectives, we thoroughly examined the statistical and machine learning methods employed in the included studies. The decision to go back twenty years was influenced by new revolutionary advancements in statistics and machine learning in producing high-quality results associated with education dropout prediction. The study closest to our systematic review was the review detailed by Chen J (2022), which looked into a few investigations estimating massive open online course (MOOC) dropouts from 2012 to 2022. Though the survey attempted to sum up the principal methods for predicting early dropout, analysis to detail the consequences of the predicted models still needs to be provided [69].

Based on our review, the development and use of predictive analysis methods that predict student dropout has been at an all-time high since 2017. Moreover, developed countries are taking a giant stride in researching the early identification of at-risk students. The quantity of articles published in the field of education dropout prediction is increasing year by year [70, 71]. Researchers still need to be satisfied with their attempts to develop modeling techniques that predict early student dropout [68]. Our review observed that the dropout expectation models were created as independent modules rather than evaluation programming in many studies. Though ensemble methods are well established to improve predictive performance [64], almost 90% of the research studies developed a model that relied on statistical or machine learning techniques. Furthermore, despite their importance, fewer models were expanded to clarify and validate the estimate of student dropout [72]. Only a few studies used advanced survival techniques and Bayesian networks for early dropout prediction. We utilized scientific classification to order the prediction models from our qualitative synthesis [70]. Random Forest, Decision Tree, Logistic Regression, Support Vector Machine, Naïve

Bayes classifier, K-Nearest Neighbour, Artificial Neural Network, and Gradient Boosting have been the most utilized methods for estimating student dropout.

The wide use of diverse statistical approaches appears across all of the included studies, which is a similarity. A wide range of techniques are used by researchers, including traditional linear models, decision-based models, and advanced ensemble methods [22, 26, 27]. This variety emphasizes that statistical methods have distinct advantages in handling the multidimensional nature of dropout prediction. In the case of evaluation metrics, to assess the effectiveness of the predictive models, accuracy and AUC were used as the most important metrics [49, 50]. Some performance metrics such as specificity and sensitivity, Kappa, F1 score, confusion matrix, absolute mean error, mean square error, C-index, and precision-recall were also used for performance evaluation [36, 48]. Numerous researches recognized the difficulty of unbalanced datasets in prediction, where dropout instances were a minority class. The necessity of methods to resolve data imbalance was highlighted in most of the studies. Still, only a few studies solved the data imbalance issue using balancing and rebalancing algorithms [47, 49, 50].

The regional focus of the investigations showed a clear difference. While some study on dropout prediction was done in developed nations, there were few studies in emerging or resource-constrained countries [44, 46, 48]. This discrepancy underlined the need for specialized solutions in various global situations. Models did not consistently consider temporal elements, such as variations in dropout risk over time. While some research focused on temporal dynamics, others mainly used static predictors [10, 21]. Studies revealed differences in the distinct types and dimensions of risk factors adopted by dropout models for prediction. Some research examined various sociodemographic, academic, and behavioral variables, while others concentrated on fewer predictors [41].

A meta-analysis was performed to determine the proportion of dropouts in the included studies to address the third objective. This investigation exposed a fundamental problem with the research on student dropout prediction related to data imbalance. Many scientific works fail to consider how dropout could be reduced in the datasets that are currently accessible. This becomes a significant issue, particularly concerning student academic performance, as dropout pupils sometimes differ from those who stay [73]. Subsequently, future research should consider developing a student dropout model that considers its data imbalance problem. Data balancing techniques have successfully addressed the issue of data imbalances in predicting student dropout using machine learning [74]. These methods have enhanced the accuracy and decreased data bias in the model. More study is required in multiple geographic areas to guarantee the adaptability of these strategies across various educational contexts. Many methods for balancing data include over-sampling, under-sampling, and combining the two. While under-sampling results in fewer occurrences of the majority class, over-sampling results in more instances of the minority class. Some data-level strategies employed are the SMOTE and Adaptive Synthetic Sampling Approach (ADASYN). Techniques at the algorithmic level, such as Balanced Random Forest and SMOTE-Bagging, have also been used [75]. A number of the included studies illustrate the use of data-balancing approaches in predicting student dropout. For example, one study that used a decision tree method with under-sampling to balance the dataset got a precision of 98.9% [29]. An AUC of 78% was obtained in a different study that combined over-sampling and under-sampling methods with a logistic classifier [49].

Additionally, several other challenges were also identified. Primarily, a large portion of every prediction technique is applied and assessed in advanced nations utilizing the available data sources gathered from developed nations. The barriers to acquiring available datasets from emerging countries necessitated the development of new datasets [76]. This includes converting the enrolment data of individuals from paper-based methodology to computerized capabilities. Moreover, to the best of academics' knowledge, only some studies have been directed at developing nations. We suggest further examination to investigate the worth of statistical learning in preventing dropouts in emerging countries.

The inability to comprehend results is one of the fundamental flaws of ML models, particularly advanced ML models, as it is challenging to ascertain the exact method that was used to mine the results. It is challenging to comprehend why a specific prediction was made in some circumstances [49]. As a result, educational managers might not believe that such models can be relied upon to support their decisions, mainly when a student's future is at risk. To promote transparency of algorithms and reliability, explainable artificial intelligence is viewed as a viable technique. This could make outputs more intelligible as well as acceptable [77, 78].

Next, most studies have only focused on general terms of early identification. To add detail, research in emerging nations must emphasize working with a more robust and exhaustive detection system that can recognize individuals in danger in forthcoming groups, grade individuals as per their likelihood of dropping from schools (getting dropped), and distinguish individuals who are in danger way before they drop [79].

Moreover, several researchers utilized academic datasets to address the issue of student dropouts. Considering the resource constraints the developing nations experience, they can use alternative methods of school-level information that may address school-related attributes and apply appropriate prediction techniques to strengthen the suggested computation analysis [80]. To advance the area of student dropout prediction, research initiatives should be expanded outside of developed countries to address the difficulties encountered by developing countries. A sophisticated strategy for preventing dropout is required because these regions frequently have diverse cultural, social, and educational backgrounds [49]. Statistical methods can be modified to meet emerging countries' distinctive needs and difficulties, resulting in more efficient and locally appropriate dropout prevention measures. Statistical models must include temporal changes. The academic process of a student is unpredictable, and dropout risks might change over time. Models that consider these temporal considerations can deliver more precise and realistic predicted outcomes [48].

Limitations

As with all research efforts, a few limitations should be acknowledged. The same goes with a wide range of surveys, and there, we may have missed some studies predicting dropout due to our selected search queries or the screening procedures.

Moreover, we focused our inquiry on the predictive models of student dropout over the last twenty years. It was also seen that a few investigations included only some exploratory and estimating characteristics, such as the dataset's quality, prediction model type, and variables affecting dropout. This, in the long run, impacts the nature of our qualitative synthesis. Sadly, many investigations did not implement a precise approach, making the evaluation more difficult. Our review was determined by the primary objective that may have outlined the review cycle, and we concluded.

Conclusion

Through a comprehensive and data-driven approach, our study delved into the crucial area of student dropout prediction, an important precursor to successful student transition. This evidence-based approach has the potential to significantly improve student outcomes and pave the way for a more successful and fulfilling educational journey. We followed Preferred Reporting Items for Systematic Reviews and Meta-Analysis and Systematic Literature Review procedures to design the survey. An overview of statistical methods for dealing with the problem of student dropout is introduced. The summary makes a few determinations; first, while a few strategies were presented for dealing with student dropout in advanced nations, there needs to be more literature about utilizing various techniques for resolving dropout prediction in emerging countries. Researchers must focus more on handling data imbalance, which is another important problem. Despite extensive efforts in employing various prediction techniques, inadequate assessment metrics are often used to evaluate model performance. Third, many experts focused on early prediction instead of positioning or even estimating components for resolving the dropout issue. Finally, when dealing with a problem, school-level datasets should be reviewed to develop alternative strategies that would assist administrators in predicting in-danger students for early intervention.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42001-023-00231-w.

Acknowledgements We thank Mr. Naman Gupta, Research Assistant, Department of Ophthalmology, Visual, and Anatomical Sciences, Wayne State University, Detroit, Michigan for his valuable support for database access throughout the process. We also thank Ms. Supriya Sathish Kumar, Research Scholar, Translational Medicine and Research, SRM Institute of Science and Technology, and reviewers for comments that greatly improved the manuscript.

Data availability All relevant data are within the manuscript and its supplementary information files.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Yukselturk, E., Ozekes, S., & Turel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and E-Learning., 17*(1), 118–133.
2. Lin, J. J. J., Imbrie ,P. K., & Reid, K. J. (2009). Student retention modelling: An evaluation of different methods and their impact on prediction results. In *2009 Research in Engineering Education Symposium REES 2009* (January).
3. Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior, 36*, 469–478.
4. Jia, P., & Maloney, T. (2015). Using predictive modelling to identify students at risk of poor university outcomes. *Higher Education, 70*(1), 127–149.
5. Chun-Teck, L. (2010). Predicting preuniversity students' mathematics achievement (published conference proceedings style). In: *International conference on mathematics education research, multimedia university*, Malaysia (pp. 299–306).
6. Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Predicting students' performance using ID3 and C4.5 classification algorithms. arXiv Preprint http://arxiv.org/abs/1310.2071
7. Durairaj, M., & Vijitha, C. (2014). Educational data mining for prediction of student performance using clustering algorithms. *International Journal of Computer Science and Information Technologies, 5*(4), 5987–5991.
8. Chen, J.-F., Hsieh, H.-N., & Do, Q. H. (2014). Predicting student academic performance: A comparison of two meta-heuristic algorithms inspired by cuckoo birds for training neural networks. *Algorithms, 7*(4), 538–553.
9. Sales, A., Balby, L., & Cajueiro, A. (2016). Exploiting academic records for predicting student drop out: A case study in Brazilian higher education. *Journal of Data, Information and Management, 7*(2), 166.
10. Ameri, S., Fard, M. J., Chinnam, R. B., & Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts. In *International conference on information and knowledge management*, 24–28 October (pp. 903–12).
11. Kitchenham, B., & Charters, S. (2007). Guidelines for performing Systematic Literature reviews in SoftwareEngineering Version 2.3. *Engineering*, *45*(4), 1051.
12. Okoli, C., & Schabram, K. (2012). A Guide to Conducting a Systematic Literature Review of Information Systems Research. *SSRN Electron J [Internet].*. https://doi.org/10.2139/ssrn.1954824.
13. Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Journal of Clinical Epidemiology, 62*(10), e1-34.
14. Moola, S. (2017). Checklist for analytical cross sectional studies. Joanna Briggs Institute Rev Man. (pp. 1–7). http://joannabriggs.org/research/critical-appraisal-tools.
15. Karmegam, D., Ramamoorthy, T., & Mappillairajan, B. (2019). A systematic review of techniques employed for determining mental health using social media in psychological surveillance during disasters. *Disaster Medicine and Public Health Preparedness, 14*(2), 265–272.
16. Lamote, C., Van Damme, J., Van Den Noortgate, W., Speybroeck, S., Boonen, T., & de Bilde, J. (2013). Dropout in secondary education: An application of a multilevel discrete-time hazard model accounting for school changes. *Quality & Quantity, 47*(5), 2425–2446.
17. Márquez-Vera, C., Romero Morales, C., & Ventura, S. S. (2013). Predicting school failure and dropout by using data mining techniques. *Revista Iberoamericana de Tecnologias del Aprendizaje, 8*(1), 7–14.
18. Şara, N. B., Halland, R., Igel, C., & Alstrup, S. (2015). High-school dropout prediction using machine learning: A Danish large-scale study. In *23rd European symposium on artificial neural networks, computational intelligence and machine learning ESANN 2015—Proc*eedings 2015 (pp. 319–324).
19. Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247–256.
20. Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. http://arxiv.org/abs/1606.06364

21. Weybright, E. H., Caldwell, L. L., Xie, H., Wegner, L., & Smith, E. A. (2017). Predicting secondary school dropout among South African adolescents: A survival analysis approach. *South African Journal of Education, 37*(2), 1–11.
22. Adelman, M., Haimovich, F., Ham, A., & Vazquez, E. (2018). Predicting school dropout with administrative data: New evidence from Guatemala and Honduras. *Education Economics, 26*(4), 356–372. https://doi.org/10.1080/09645292.2018.1433127
23. Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2021). Early detection of students at risk—Predicting student dropouts using administrative student data and machine learning methods. *SSRN Electronic Journal, 11*(3), 1–41.
24. Hegde, V., & Prageeth, P. P. (2018). Higher education student dropout prediction and analysis through educational data mining. In: *2018 2nd international conference on inventive systems and control (ICISC)*. IEEE [cited 2021 Oct 14]. https://ieeexplore.ieee.org/document/8398887/
25. Nagy, M., & Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. In *2018 IEEE 22nd international conference on intelligent engineering systems (INES)*. IEEE [cited 2021 Oct 14]. https://ieeexplore.ieee.org/document/8523888/
26. Lee S, Chung JY. The machine learning-based dropout early warning system for improving the performance of dropout prediction. Appl Sci. 2019;9(15).
27. da Silva, P. M., Lima, M. N. C. A., Soares, W. L., Silva, I. R. R., de Fagundes, R. A. A., de Souza, F. F. (2019). Ensemble regression models applied to dropout in higher education. In *2019 8th Brazilian conference on intelligent systems (BRACIS)*. IEEE [cited 2021 Oct 14]. https://ieeexplore.ieee.org/document/8923655/
28. Al-Shabandar, R., Hussain, A. J., Liatsis, P., & Keight, R. (2019). Detecting at-risk students with early interventions using machine learning techniques. *IEEE Access., 7*, 149464–149478.
29. Barros, T. M., Neto, P. A. S., Silva, I., & Guedes, L. A. (2019). Predictive models for imbalanced data: A school dropout perspective. *Education Sciences, 9*(4), 275.
30. Biswas, A. A., Majumder, A., Mia, M. J., Nowrin, I., & Ritu, N. A. (2019). Predicting the enrollment and dropout of students in the post-graduation degree using machine learning classifier. *International Journal of Innovative Technology and Exploring Engineering, 8*(11), 3083–3088.
31. Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. Vol. 12163 *LNAI, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer International Publishing (pp. 129–140). https://doi.org/10.1007/978-3-030-52237-7_11
32. Tenpipat, W., & Akkarajitsakul, K. (2020). Student dropout prediction: A KMUTT case study. In *2020 1st international conference on big data analytics and practices (IBDAP)*. IEEE [cited 2021 Oct 14]. https://ieeexplore.ieee.org/document/9245457/
33. Patacsil, F. F. (2020). Survival analysis approach for early prediction of student dropout using enrollment student data and ensemble models. *Universal Journal of Educational Research, 8*(9), 4036–4047.
34. Baranyi, M., Nagy, M., & Molontay, R. (2020). Interpretable deep learning for university dropout prediction. In *SIGITE 2020—Proceedings 21st annual conference on information technology education* (pp. 13–9).
35. Nangia, S., Anurag, J., & Gambhir, I. (2020). A machine learning approach to identity the students at the risk of dropping out of secondary education in India. *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-981-15-2475-2_51
36. Lottering, R., Hans, R., & Lall, M. (2020). A machine learning approach to identifying students at risk of dropout: A case study. *International Journal of Advanced Computer Science and Applications, 11*(10), 417–422.
37. Freitas, F. A. D., Vasconcelos, F. F. X., Peixoto, S. A., Hassan, M. M., Dewan, M. A. A., de Albuquerque, V. H. C., et al. (2020). IoT system for school dropout prediction using machine learning techniques based on socioeconomic data. *Electronics, 9*(10), 1613.
38. Maldonado, S., Miranda, J., Olaya, D., Vásquez, J., & Verbeke, W. (2021). Redefining profit metrics for boosting student retention in higher education. *Decision Support Systems, 143*(August 2020), 113493.
39. Opazo, D., Moreno, S., Álvarez-Miranda, E., & Pereira, J. (2021). Analysis of first-year university student dropout through machine learning models: A comparison between universities. *Mathematics., 9*(20), 1–27.

40. Fernandez-Garcia, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sanchez-Figueroa, F. (2021). A real-life machine learning experience for predicting university drop-out at different stages using academic data. *IEEE Access., 9*, 133076–133090.
41. Queiroga, E. M., Batista Machado, M. F., Paragarino, V. R., Primo, T. T., & Cechinel, C. (2022). Early prediction of at-risk students in secondary education: A countrywide K-12 learning analytics initiative in Uruguay. *Information, 13*(9), 1–25.
42. Segura, M., Mello, J., & Hernandez, A. (2022). Machine learning prediction of university student dropout: Does preference play a key role? *Mathematics., 10*(18), 3359.
43. Moreira da Silva, D. E., Solteiro Pires, E. J., Reis, A., de Moura Oliveira, P. B., & Barroso, J. (2022). Forecasting students dropout: A UTAD university study. *Future Internet., 14*(3), 1–14.
44. Mnyawami, Y. N., Maziku, H. H., & Mushi, J. C. (2022). Comparative study of AutoML approach, conventional ensemble learning method, and KNearest Oracle-AutoML model for predicting student dropouts in Sub-Saharan African countries. *Applied Artificial Intelligence, 36*(1), 2145632.
45. Dake, D. K., & Buabeng-Andoh, C. (2022). Using machine learning techniques to predict learner drop-out rate in higher educational institutions. *Mobile Information Systems, 2022*, 1–9.
46. Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence., 3*(March), 100066. https://doi.org/10.1016/j.caeai.2022.100066
47. Flores, V., Heras, S., & Julian, V. (2022). Comparison of predictive models with balanced classes using the SMOTE method for the forecast of student dropout in higher education. *Electronics, 11*(3), 457.
48. Garcia-Zanabria, G., Gutierrez-Pachas, D. A., Camara-Chavez, G., Poco, J., & Gomez-Nieto, E. (2022). SDA-Vis: A visualization system for student dropout analysis based on counterfactual exploration. *Applied Sciences, 12*(12), 5785.
49. Selim, K. S., & Rezk, S. S. (2023). On predicting school dropouts in Egypt: A machine learning approach. *Education and Information Technologies, 28*, 9235–9266. https://doi.org/10.1007/s10639-022-11571-x
50. Song, Z. H., Sung, S. H., Park, D., & Park, B. K. (2023). All-year dropout prediction modeling and analysis for university students. *Applied Sciences, 13*(2), 1143.
51. Wang, N. (2016). *How to conduct a meta-analysis of proportions in R: A comprehensive tutorial*. John Jay College Criminal Justice (June):1–63.
52. Zheng, Z., Cai, Y., & Li, Y. (2015). Oversampling method for imbalanced classification. *Computer Informatics., 34*(5), 1017–1037.
53. Learned-Miller, E. G. (2014). *Introduction to Supervised Learning* (p. 3). Amherst, MA, USA: Department of Computer Science, University of Massachusetts. https://people.cs.umass.edu/~elm/Teaching/Docs/supervised2014a.pdf
54. Kumar, M., Singh, A. J., & Handa, D. (2017). Literature survey on educational dropout prediction. *International Journal of Education and Management Engineering, 7*(2), 8.
55. Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science., 72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157
56. Gray, G., McGuinness, C., & Owende, P. (2014). An application of classification models to predict learner progression in tertiary education. In *2014 IEEE international advance computing conference (IACC)*. IEEE (pp. 549–554).
57. Arsad, P. M., & Buniyamin, N. (2013). A neural network students' performance prediction model (NNSPPM). In *2013 IEEE international conference on smart instrumentation, measurement and applications (ICSIMA)*. IEEE (pp. 1–5).
58. Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence, 2*(2), 34–38.
59. 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.
60. Kartal, O. O. (2015). *Using survival analysis to investigate the persistence of students in an introductory information technology course at METU*. Middle East Technical University.
61. Li, Y., Yang, T., Zhou, J., & Ye, J. (2018). A multi-task learning formulation for survival analysis. *Proceedings of SIGKDD*. https://doi.org/10.1137/1.9781611975321.33

62. Bani, M. J., & Haji, M. (2017). College student retention: When do we losing them? arXiv Preprint http://arxiv.org/abs/1707.06210
63. Zohair, L. M. A. (2019). Prediction of student's performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education, 16*(1), 1–18.
64. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. Predicting academic performance: A systematic literature review. In *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education* (pp. 175–99).
65. Kaliannan, M., & Chandran, S. D. (2012). Empowering Students through Outcome-Based Education (OBE). *Res Educ[Internet].*, *87*(1), 50–63.
66. Arroway, P., Morgan, G., O'Keefe, M., & Yanosky, R. (2016). *Learning analytics in higher education*. Research report. ECAR, Louisville, CO.
67. Rajak, A., Shrivastava, A. K., & Shrivastava, D. P. (2018). Automating outcome based education for the attainment of course and program outcomes. In *2018 Fifth HCT Information Technology Trends (ITT)*. IEEE (pp. 373–376).
68. Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences, 11*(1), 1–28.
69. Chen, J., Fang, B., Zhang, H., & Xue, X. (2022). A systematic review for MOOC dropout prediction from the perspective of machine learning. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2022.2124425
70. Manjarres, A. V., Sandoval, L. G. M., & Suárez, M. S. (2018). Data mining techniques applied in educational environments: Literature review. *Digital Education Review, 33*, 235–266.
71. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355.
72. Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310.
73. Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications, 41*(2), 321–330.
74. Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys, 52*(4), 1–36.
75. Mduma, N. (2023). Data balancing techniques for predicting student dropout using machine learning. *Data, 8*(3), 49.
76. Mgala, M., & Mbogho, A. (2015). Data-driven intervention-level prediction modeling for academic performance. In *Proceedings of the seventh international conference on information and communication technologies and development* (pp. 1–8).
77. Adadi, A., & Berrada, M. (2020). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access., 2018*(6), 52138–52160.
78. Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). *Education and Information Technologies, 28*, 8299–8333. https://doi.org/10.1007/s10639-022-11536-0
79. Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1909–1918).
80. Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal, 18*(1), 1–10.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Raghul Gandhi Venkatesan[1] · **Dhivya Karmegam**[2] · **Bagavandas Mappillairaju**[3]

✉ Bagavandas Mappillairaju
bagwandm@srmist.edu.in; mbdas49@gmail.com

Raghul Gandhi Venkatesan
rv4032@srmist.edu.in; raghulgandhivenkatesan@gmail.com

Dhivya Karmegam
dhivya.megam@gmail.com

[1] Department of Mathematics, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu 603203, India

[2] School of Public Health, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu 603203, India

[3] Centre for Statistics, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu 603203, India