Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 19/Jul/2023; Camera ready: 20/May/2024; 1st round notif.: 02/Jan/2024; Edition review: 20/May/2024; New version: 05/Jan/2024; Available online: 20/May/2024; 2nd round notif.: 04/Mar/2024; Published: 20/May/2024;

Educational Data Mining for Dropout Prediction: Trends, Opportunities, and Challenges

Miriam Pizzatto Colpo
Programa de Pós-Graduação em Computação
Universidade Federal de Pelotas (UFPel)
Diretoria de Tecnologia da Informação
Instituto Federal Farroupilha (IFFar)
ORCID: 0000-0002-6477-3227
miriam.colpo@inf.ufpel.edu.br

Tiago Thompsen Primo
Programa de Pós-Graduação em Computação
Universidade Federal de Pelotas (UFPel)
ORCID: 0000-0003-3870-097X
tiago.primo@inf.ufpel.edu.br

Marilton Sanchotene de Aguiar
Programa de Pós-Graduação em Computação
Universidade Federal de Pelotas (UFPel)
ORCID: 0000-0002-5247-6022
marilton@inf.ufpel.edu.br

Cristian Cechinel
Programa de Pós-Graduação em Computação
Universidade Federal de Pelotas (UFPel)
Centro de Ciências, Tecnologias e Saúde
Universidade Federal de Santa Catarina (UFSC)
ORCID: 0000-0001-6384-409X
contato@cristiancechinel.pro.br

Abstract

Today, we face academic, social, and economic losses associated with student dropouts. Several studies have applied data mining techniques to educational datasets to understand dropout profiles and recognize at-risk students. To identify the contextual (academic levels, modalities, and systems), technical (tasks, categories of algorithms, and tools), and data (types, coverage, and volume) characteristics related to these works, we performed a systematic literature review, considering institutional and academic degree dropout. Internationally recognized repositories were searched, and the selected articles demonstrated, among other characteristics, a greater exploration of educational, demographic, and economic data of undergraduate students from classification techniques of decision tree ensembles. In addition to not having identified any study from underdeveloped countries among the selected ones, we found shortcomings in the application of predictive models and in making their predictions available to academic managers, which suggests an underutilization of the efforts and potential of most of these studies in educational practice.

Keywords: *Student Dropout; Dropout Prediction; Educational Data Mining; Systematic Literature Review.*

Cite as: Colpo, M. P., Thompsen Primo, T., Aguiar, M. S. & Cechinel, C. (2024). Educational Data Mining for Dropout Prediction: Trends, Opportunities, and Challenges. Revista Brasileira de Informática na Educação, 32, 220-256. https://doi.org/10.5753/rbie.2024.3559.

1 Introduction

During the COVID-19 pandemic and the need for social distancing, the formal education, until then on-site, changed to remote and virtual classes. Although essential to the pandemic, the emergency remote education posed challenges to students, professors, and institutions, which needed to be methodologically and structurally prepared for this alternative form of teaching (J. R. Santos & Zaboroski, 2020). In this scenario of difficulties and changes, student dropout caused an even more significant concern (Colpo et al., 2021).

The student dropout, characterized by the student's permanent departure from your degree of origin without proper completion (Brasil, 1996), represents a severe educational problem that causes academic, social, and economic damage to the individual who evades and to society. This fact is because there is a correlation between the education level and the salary gains of the population, which directly affects the socio-economic development of a country (Pontili et al., 2018). In addition, shortages of skilled labor can negatively affect the productive capacity of a nation, and students who have not completed their degrees are more likely to be unemployed and need welfare benefits (Lee & Chung, 2019). When considering public education, student dropout also means a waste of state resources, not only financial but also personnel and infrastructure (Silva Filho et al., 2007).

Identifying at-risk students, based on knowledge of the previous dropout patterns, can support educational institutions in the decision-making and implementing institutional policies to prevent dropout (Nagy & Molontay, 2018). With this purpose, some studies try to draw dropout profiles from manual analyses based on data collected from interviews or surveys. However, because of the difficulty of contacting a significant portion of the evaded population, the results are susceptible to biases (Silva, 2013). Moreover, the large volumes of educational data provided by learning and academic management systems also make manual explorations unfeasible, although they store information of high strategic potential (Romero & Ventura, 2020).

In this context, Educational Data Mining (EDM) techniques have been adopted to automate dropout analysis. EDM is a research area that focuses on developing methods capable of exploring large volumes of educational data to understand more effectively the students' behavior and other factors related to learning (Baker et al., 2011). As a multidisciplinary area, data mining (academic or not) incorporates techniques from different domains, such as machine learning (Han et al., 2012). Supervised machine learning, for example, is the basis of classification, which is the EDM task often used in dropout prediction for automatically discovering patterns and attributes related to this phenomenon and for automatically identifying at-risk students. In the first scenario, classification models are developed descriptively to improve the understanding of dropout-related patterns and help professionals identify at-risk students through pedagogical monitoring (manual prediction). In the second one, the models are intended to identify potential dropouts (automatic prediction) (Colpo et al., 2020).

Several secondary studies have already been conducted on applying data mining and machine learning techniques to educational problems. Specifically considering the context of student dropout, Mduma et al. (2019a) presented a survey on machine learning techniques used to predict dropout. Marques et al. (2019) and Rondado de Sousa et al. (2021) carried out systematic mappings on data mining technologies in identifying the causes of dropout and addressing the problem of face-to-face students dropout, respectively. Unlike Mduma et al. (2019a), Marques et al. (2019) and Rondado de Sousa et al. (2021), in this work, we analyze studies that apply EDM techniques in the dropout prediction context through a Systematic Literature Review (SLR). In addition, the scope of this research is broader than that considered by Marques et al. (2019) since it is not restricted to identifying the causes of dropout, and our RSL includes contextual, technical, and data aspects not covered by Mduma et al. (2019a) and Marques et al. (2019). SLRs also were conducted by Colpo et al. (2020), de Oliveira et al. (2021), and Agrusti et al. (2019). The first study focused on the Brazilian research scenario, while the latter two concentrated on university dropouts. In contrast, our RSL covers the international research scene without restrictions on the level or modality of education. However, we have limited our scope to studies that address the problem of breaking the link with formal education.

In more detail, this SLR analyzes studies that use EDM to predict institutional and degree dropout. The aim is to identify (*i*) contextual characteristics, including educational modalities, levels, and systems; (*ii*) technical characteristics as tasks, categories of algorithms, and tools; and (*iii*) data characteristics considering types, coverage, and volume. We do not view works that deal with dropout in subjects/courses, as it does not reflect the closure of students' educational bonds.

The planning of this SLR, including the method and definition of research questions, search strategies, and the selection and quality criteria, was described in Section 2. At the same time, we explained the details of its conduct in Section 3. In Section 4, we present the results obtained, answer previously established research questions, and describe many of the studies analyzed. Then, in section5, we summarize some trends, opportunities, and challenges that we have observed among the studies that apply EDM in predicting student dropout, as well as possible threats to the validity of our research. Last, conclusions and final remarks are presented in Section 6.

2 Methodology and Planning

According to Kitchenham and Charters (2007), developing an SLR is identifying and evaluating a broad range of research in an area or topic of interest using a suitable and reliable method. In this paper, we used the SLR method proposed by Kitchenham and Charters (2007) to investigate the international research scene regarding the use of EDM in predicting student dropout. More specifically, as the primary research question, we sought to examine these studies' contextual, technical, and data characteristics, considering dropout at the institutional and degree levels/scopes. To this end, we believe the following specific research questions in this SLR:

- Q1. What are the objectives of these studies?
- Q2. What educational levels, modalities, and systems were investigated?
- Q3. What is the nature, coverage, and volume of the data used?
- Q4. What tasks, techniques, and tools are being focused on?

This SLR considered works published from 2016 to 2022 in the following scientific bases: ACM Digital Library (https://dl.acm.org), IEEE Xplore (https://ieeexplore.ieee.org), Web of Science (https://www.webofscience.com), and Scopus (https://www.scopus.com). We contemplated publications from the last seven years before the execution of the SLR since we consider this sample to be reasonably representative of the advances and the current research scenario. In addition, we chose the scientific databases based on their international relevance to Informatics in the Education community.

The expressions used in the search engines were previously refined through the execution and evaluation of preliminary queries. They included terms related to the Problem ("drop*out" OR "dropout"), Population ("student" OR "school" OR "college" OR "university”), and Intervention ("data mining" OR "machine learning"). However, as the repositories use different default search forms, we have standardized that the query should be performed on the publication’s metadata (title, abstract, and keywords). We achieved this by creating advanced queries adapted to each repository’s different options and syntaxes, as shown in Table 1.

Table 1: Advanced search expressions for each scientific database.

Database	Advanced Search Expressions
ACM	Abstract:(("drop*out" OR "dropout") AND ("student" OR "school" OR "college" OR "university") AND ("data mining" OR "machine learning")) OR Title:(("drop*out" OR "dropout") AND ("student" OR "school" OR "college" OR "university") AND ("data mining" OR "machine learning")) OR Keyword:(("drop*out" OR "dropout") AND ("student" OR "school" OR "college" OR "university") AND ("data mining" OR "machine learning"))
IEEE	(("All Metadata": "drop*out" OR "All Metadata": "dropout") AND ("All Metadata": "student" OR "All Metadata": "school" OR "All Metadata": "college" OR "All Metadata": "university") AND ("All Metadata": "data mining" OR "All Metadata": "machine learning"))
Scopus	TITLE-ABS-KEY(("drop*out" OR "dropout") AND ("student" OR "school" OR "college" OR "university") AND ("data mining" OR "machine learning"))
WoS	TS=(("drop*out" OR "dropout") AND ("student" OR "school" OR "college" OR "university") AND ("data mining" OR "machine learning"))

For the selection of studies, we established five criteria:

- the article must be written in Portuguese[1] or English, following the format of a scientific paper;
- the study should describe the application of a specific solution, thus excluding literature reviews;
- the article must adopt EDM techniques in solutions that contribute to the prediction of student dropout;
- the work should address dropout in an institutional or academic degree scope, discarding research aimed at dropout in subjects/courses or short-term training; and

[1] Although we take into account the international research scene, we kept the Portuguese language among the acceptance/selection criteria because it is also in our domain and because we felt that this decision would not significantly influence the results, considering that the searches were carried out with keywords in English and international scientific databases. This judgment was confirmed during the RSL process, as only two of the selected studies (to be presented in the next section) are written in Portuguese.

- The article must provide details of its development and results, transposing the proposal sphere.

Finally, to refine the selection/exclusion process, we also observed three quality criteria:

- the paper should be more than five pages long (whether single or double column) to present a minimum of detail about its solution;
- in its validations, the research must consider a dataset of at least 500 instances/students to guarantee a minimum representativeness/quality to the experiments; and
- The data used in the research should not be collected mostly by questionnaires and surveys due to the higher propensity of bias in this approach.

3 Conducting the RSL

We conducted the RSL in 2023 with the support of the Parsifal tool[2], considering three phases. Initially, in **Step 1**, we query the scientific databases[3] and export their results' metadata (including abstracts) in BibTeX format. Then, we import the BibTeX files into Parsifal to expedite the duplicate check and streamline the article review process. After eliminating exact duplicates, in **Step 2**, the studies resulting from Step 1 had their abstracts and metadata (such as size, format, and language) analyzed, and papers that did not meet the previously established selection or quality criteria got removed from the analysis stage. Note that although the abstract is present, not all publication metadata includes pages, language, and format information. Consequently, some articles that may have been excluded at this stage proceeded to the next phase. It is also important to point out that we did not conduct a peer review in the analysis of the publications due to the large volume of articles collected. That is, a single researcher analyzed each article. Finally, in **Step 3**, we extracted the full texts of the studies resulting from Step 2, when available[4], and analyze them based on our selection and quality criteria. Furthermore, we only retained the most recent studies when detecting similar publications by the same authors in our reading. These articles represent progress in the same research, even though they are not identical duplicates.

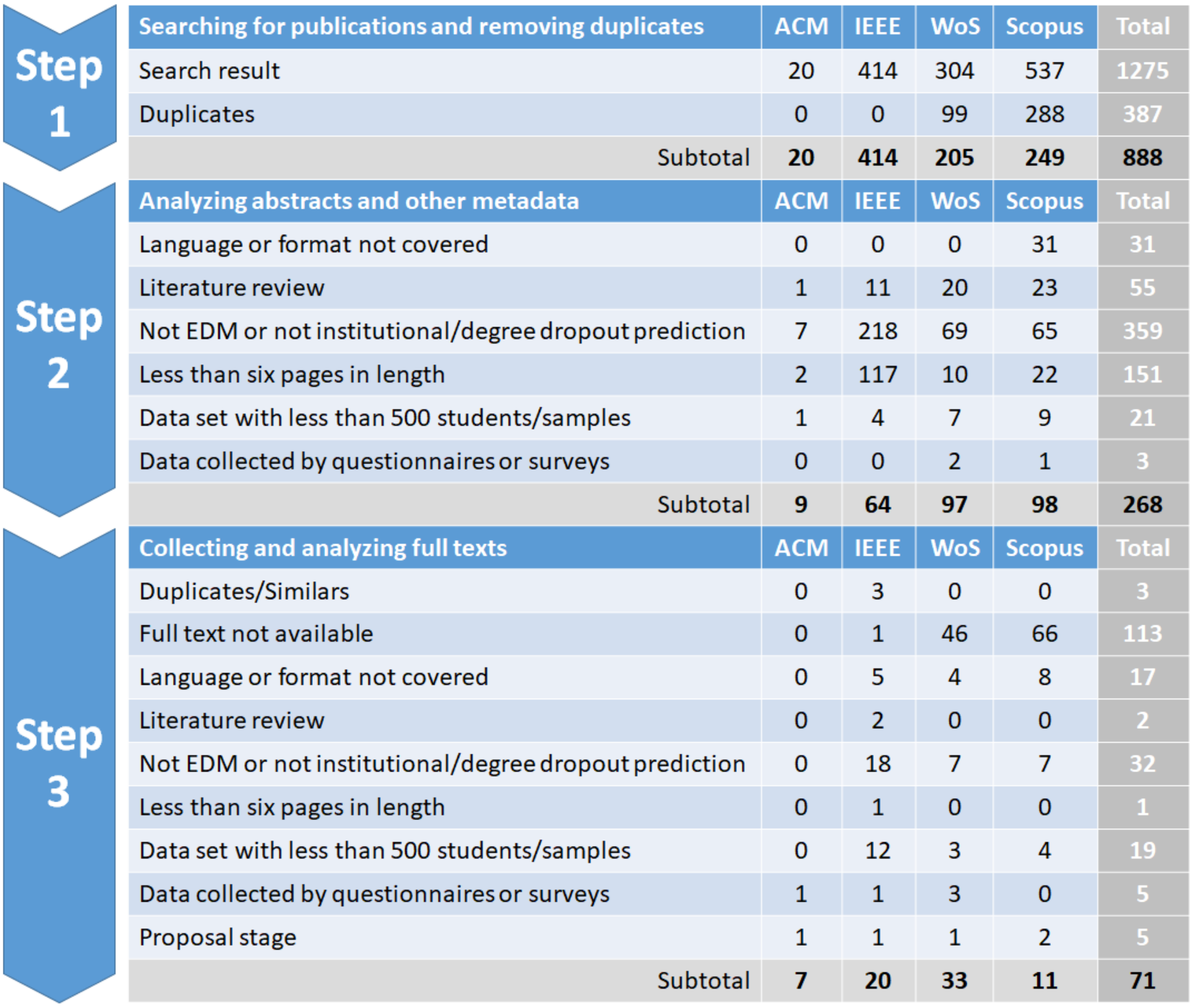
Figure 1 shows, by repository, the number of results returned in the searches and the number of exclusions for each criterion analyzed in each step. Notice that in Step 1, the searches yielded 1275 results, and many of the studies returned by the WoS (99) and Scopus (288) databases were identified as duplicates and removed. This is because these databases were the last to be searched, and the other repositories had already retrieved many of their results. It is also possible to observe that during the analysis of the abstracts in Step 2, many studies were removed because they did not use EDM and machine learning techniques or because they did not address the problem of

[2] https://www.parsif.al/

[3] The scientific databases were accessed through the Portal de Periódicos (https://www.periodicos.capes.gov.br) of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), a foundation linked to the Brazilian Ministry of Education.

[4] Notice that the availability of the full text is not among the selection criteria presented in Section 2. This restriction is a standard collection criterion since selection depends on the feasibility of the analysis. The same applies to the duplicate limitation, as there is no reason to analyze an article that is identical/similar to another.

institutional or degree dropout. More specifically, many studies dealt with dropouts from courses/subjects, Massive Open Online Courses (MOOCs), or other educational problems such as grade prediction. Furthermore, especially considering the IEEE database, many articles were entirely outside the educational domain and related to the use of the dropout technique in the regularization of neural network models. Finally, in Step 3, the most significant discards were publications whose files/full texts were not available/accessible. Since the ACM and IEEE Digital Libraries mostly hold their publications, these occurrences were concentrated in the WoS and Scopus databases.



	Searching for publications and removing duplicates	ACM	IEEE	WoS	Scopus	Total
Step 1	Search result	20	414	304	537	1275
	Duplicates	0	0	99	288	387
	Subtotal	**20**	**414**	**205**	**249**	**888**
	Analyzing abstracts and other metadata	**ACM**	**IEEE**	**WoS**	**Scopus**	**Total**
Step 2	Language or format not covered	0	0	0	31	31
	Literature review	1	11	20	23	55
	Not EDM or not institutional/degree dropout prediction	7	218	69	65	359
	Less than six pages in length	2	117	10	22	151
	Data set with less than 500 students/samples	1	4	7	9	21
	Data collected by questionnaires or surveys	0	0	2	1	3
	Subtotal	**9**	**64**	**97**	**98**	**268**
	Collecting and analyzing full texts	**ACM**	**IEEE**	**WoS**	**Scopus**	**Total**
Step 3	Duplicates/Similars	0	3	0	0	3
	Full text not available	0	1	46	66	113
	Language or format not covered	0	5	4	8	17
	Literature review	0	2	0	0	2
	Not EDM or not institutional/degree dropout prediction	0	18	7	7	32
	Less than six pages in length	0	1	0	0	1
	Data set with less than 500 students/samples	0	12	3	4	19
	Data collected by questionnaires or surveys	1	1	3	0	5
	Proposal stage	1	1	1	2	5
	Subtotal	**7**	**20**	**33**	**11**	**71**

Figure 1: Progress of the article analysis steps, taking into account the selection and quality criteria specified in the RSL protocol.

Figure 2(a) graphically shows the number of articles resulting from each selection phase, considering the scientific databases separately. We selected 71 works identified in Tables 2 and 3. Figure 2(b) shows the distribution of the selected studies by year of publication. It can be seen that no article published in 2016 was selected and that there has been a progressive increase in the number of selected papers since 2019. These results suggest a recent growth in the interest to apply EDM techniques to face the student dropout problem, confirming the potential of this research topic.

Regarding the origin of publications, Figure 3 shows the mapping of the number of selected

Table 2: List of selected articles.

#	Reference	Country(ies) of Affiliations	Database
1	Baranyi et al. (2020)	Hungary	ACM
2	Chen et al. (2018)	USA	ACM
3	Kang and Wang (2018)	USA	ACM
4	Shiau (2020)	China	ACM
5	Vasquez Verdugo et al. (2022)	Chile, USA	ACM
6	Xu and Wilson (2021)	USA	ACM
7	Yu et al. (2021)	USA	ACM
8	Aguirre and Pérez (2020)	Ecuador, Spain	IEEE
9	Bassetti et al. (2022)	Italy	IEEE
10	Böttcher et al. (2021)	Germany	IEEE
11	Costa et al. (2021)	Brazil, Chile	IEEE
12	da Silva et al. (2019)	Brazil	IEEE
13	Fernández-García et al. (2021)	Spain	IEEE
14	Kiss et al. (2019)	Hungary	IEEE
15	Kurniawati and Maulidevi (2022)	Indonesia	IEEE
16	Lottering et al. (2020)	South Africa	IEEE
17	Masood and Begum (2022)	India	IEEE
18	Mduma and Machuve (2021)	Tanzania	IEEE
19	Nagy and Molontay (2018)	Hungary	IEEE
20	Orooji and Chen (2019)	USA	IEEE
21	Ortigosa et al. (2019)	Spain	IEEE
22	Pachas et al. (2021)	Brazil, Peru	IEEE
23	Prada et al. (2020)	Italy, Poland, Portugal, Romania, Spain	IEEE
24	G. Santos et al. (2020)	Brazil	IEEE
25	Solis et al. (2018)	Costa Rica	IEEE
26	Yang et al. (2021)	USA	IEEE
27	Yoo et al. (2017)	USA	IEEE
28	Alturki et al. (2022)	Germany	WoS
29	Barros et al. (2019)	Brazil	WoS
30	Beaulac and Rosenthal (2019)	Canada	WoS
31	Berka and Marek (2021)	Czech Republic	WoS
32	Chung and Lee (2019)	South Korea, USA	WoS
33	Crespo (2020)	United Kingdom	WoS
34	de Assis et al. (2022)	Brazil	WoS
35	Deho et al. (2022)	Australia	WoS
36	Del Bonifro et al. (2020)	Italy, France	WoS
37	Demeter et al. (2022)	USA	WoS
38	Flores et al. (2022)	Peru, Spain	WoS
39	Fontana et al. (2021)	Italy	WoS
40	Freitas et al. (2020)	Brazil, Saudi Arabia, Canada	WoS
41	Hannaford et al. (2021)	USA	WoS
42	Hoffait and Schyns (2017)	Belgium	WoS
43	Iam-On and Boongoen (2017a)	Thailand	WoS
44	Iam-On and Boongoen (2017b)	Thailand	WoS
45	Karimi-Haghighi et al. (2022)	Spain	WoS

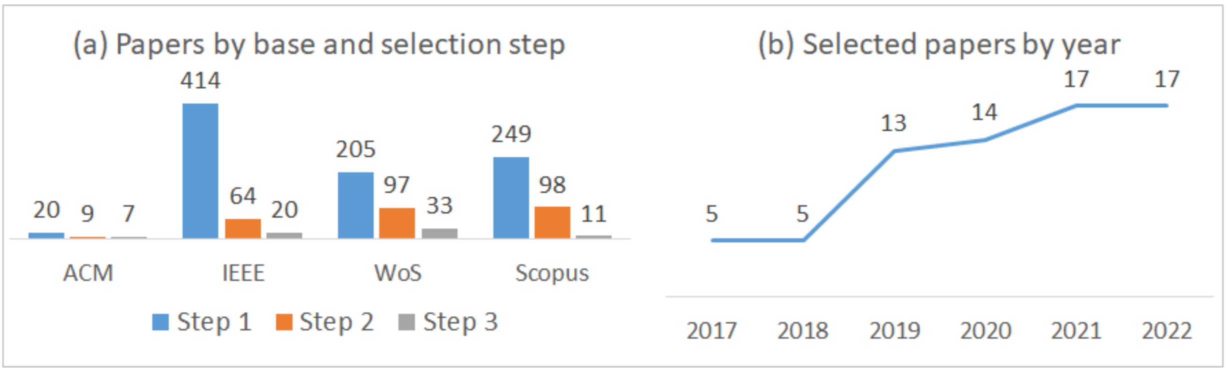


Figure 2: Status by database and step (a); and temporal evolution of selected articles (b).

Table 3: List of selected articles (continuation of Table 2).

#	Reference	Country(ies) of Affiliations	Database
46	Kuzilek et al. (2021)	Czech Republic, Germany	WoS
47	Lee and Chung (2019)	USA, South Korea	WoS
48	Opazo et al. (2021)	Chile	WoS
49	Palacios et al. (2021)	Chile	WoS
50	Perchinunno et al. (2021)	Italy	WoS
51	Perez et al. (2018)	Colombia	WoS
52	Queiroga et al. (2022)	Brazil, Uruguay	WoS
53	Queiroga et al. (2020)	Brazil, Chile	WoS
54	Segura et al. (2022)	Spain, Paraguay	WoS
55	Shilbayeh and Abonamah (2021)	United Arab Emirates	WoS
56	Sorensen (2019)	USA	WoS
57	Tsai et al. (2020)	Taiwan	WoS
58	Urbina-Najera and Mendez-Ortega (2022)	Mexico	WoS
59	Villegas-Ch et al. (2020)	Ecuador, Spain	WoS
60	Viloria et al. (2019)	Colombia	WoS
61	Agrusti et al. (2020)	Italy	Scopus
62	Bitencourt et al. (2022)	Brazil	Scopus
63	Gamao and Gerardo (2019)	Philippines	Scopus
64	Hutagaol and Suharjito (2019)	Indonesia	Scopus
65	Mduma et al. (2019b)	Tanzania	Scopus
66	Naseem et al. (2022)	Fiji	Scopus
67	Nuanmeesri et al. (2022)	Thailand	Scopus
68	Oreshin et al. (2020)	Russia	Scopus
69	Park and Yoo (2021)	South Korea	Scopus
70	Rovira et al. (2017)	Spain	Scopus
71	Vega et al. (2022)	Peru	Scopus

studies by country, considering the affiliations of the authors of each article. For reference, the graphic also presents the continent, the Gross National Income (GNI) per capita (gray line), and the Human Development Index – HDI (blue line)[5] for each country. Although the HDI considers

[5] The HDI aims to assess a country's development based on the capabilities of its population, not just its economic growth. Therefore, it considers three critical dimensions of human development: income, education, and health.

multiple dimensions, it is possible to observe an alignment between the two indicators so that a country with a lower GNI also tends to have a lower HDI. A greater concentration of works is observed in Europe and South and North America[6], with the United States (USA), Brazil, and Spain standing out as the countries with the most considerable quantities of publications. Although most studies have been conducted in developed countries with higher GNI and very high HDI (equal to or greater than 0.8), the interest in developing countries with high HDI (between 0.7 and 0.8) in this area of research is also notable, especially in South America. However, only two and one of the selected articles concern countries with medium (HDI below 0.7) and low (HDI below 0.55) levels of human development, respectively, and with lower GNIs. This fact shows that the growth of research aimed at applying EDM to the problem of student dropout is not confirmed in the poorest and underdeveloped countries, precisely those that would most need this type of initiative. As Section 1 mentions, reducing student dropout is essential for socioeconomic development. Therefore, there is still a need to publish and disseminate knowledge on this topic to instigate the development of new studies, especially in countries lacking it.

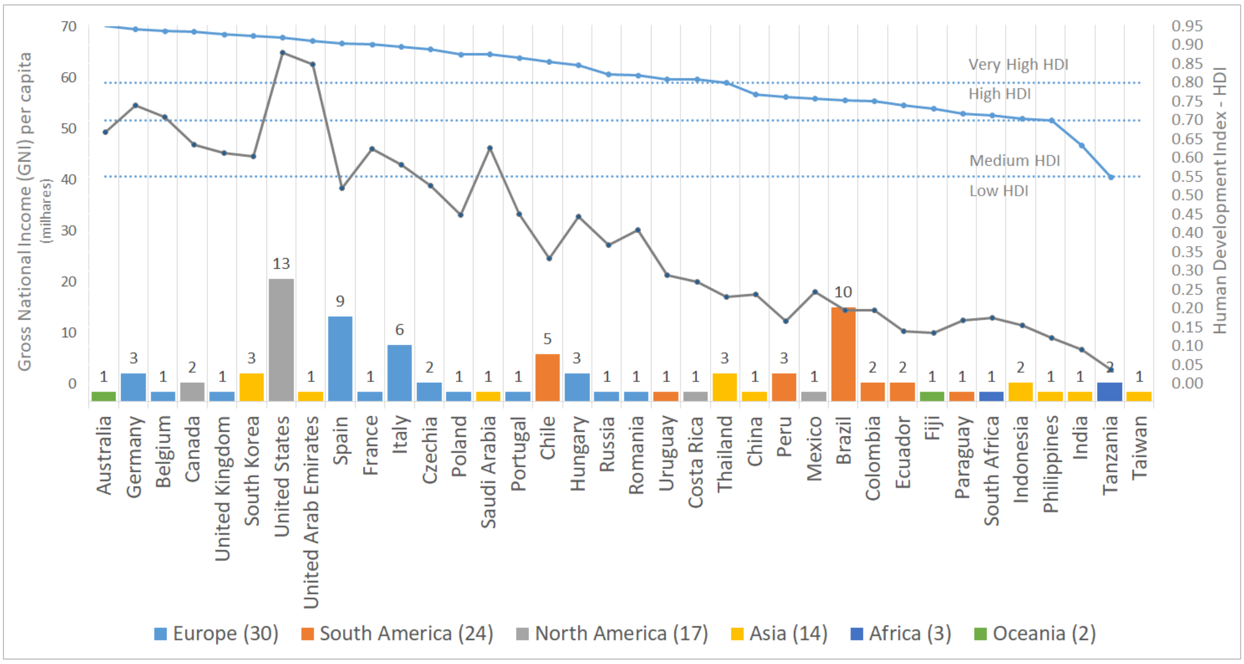


Figure 3: Distribution of selected articles by country and continent.

4 Results and Discussion

At the end of the selection process, we extracted information from the selected articles to answer the research questions established in Section 2. The results were synthesized and discussed in this section, considering the following organization: Subsection 4.1 will present the results from

The GNI per capita and HDI values shown in Figure 3 were taken from the last HDI dataset, available in December 2023 at https://hdr.undp.org/data-center/human-development-index#/indicies/HDI. Notice that the Taiwan/Republic of China does not have GNI and HDI information in Figure 3, as the United Nations does not recognize it.

[6] Notice that the total number of articles from each continent is presented with its respective legend, in parentheses.

contextual perspectives related to Q1 and Q2. Subsections 4.2 and 4.3 will describe characteristics of the data and techniques used, answering Q3 and Q4, respectively.

4.1 Context

4.1.1 Objectives

Regarding the first research question, the works that use EDM techniques in the context of student dropout, in general, aim at (*i*) the automatic discovery of patterns and attributes related to this phenomenon to understand its causes better and thus help professionals to predict at-risk students through pedagogical monitoring ("manual" prediction); (*ii*) the development of predictive models that can automatically identify students with dropout patterns (automatic prediction); or (*iii*) both of the above objectives, that is, they seek to provide predictive models of dropout, in addition to investigating patterns or indications of the attributes most strongly associated with dropout. Figure 4 presents the distribution of the articles selected in this RSL concerning these objectives. Notice that only six studies (8%) are restricted to investigating the patterns and relationships between the variables (attributes) analyzed and student dropout. The vast majority, 65 (92%), focus on developing predictive models, among which a large proportion, 46 (65% of the total number of articles), also address the investigation of patterns or attributes related to dropout.

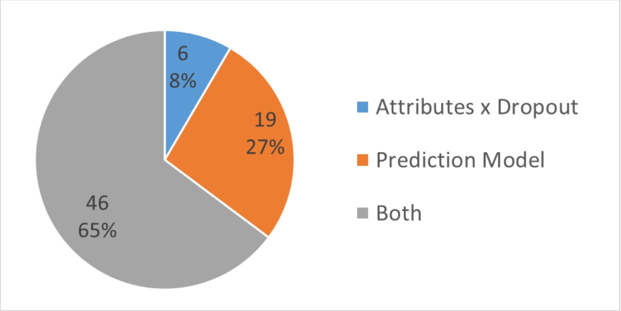


Figure 4: Distribution of the selected articles according to their objectives.

Considering the first category, in common, the six studies demonstrate a strong relationship between student dropout and students' low academic performance. In more detail, Iam-On and Boongoen (2017a) used the clustering approach to build descriptive models related to dropout, considering demographic/social and academic (school and university) information from students at Mae Fah Luang University (Thailand). Groups were generated to evaluate newly admitted students and those who completed the first year of their course/degree separately. From the analysis of the representative profile of each group, the authors identified that students with high academic performance in school continue to perform well in university and that students with low academic performance tend to drop out after the first year.

Yoo et al. (2017) and de Assis et al. (2022) used academic and demographic, and only academic data in association rule mining to identify patterns related to the dropout of Computer Science students from a public university in the USA and of Production Engineering students at the Federal Center for Technological Education of Rio de Janeiro (Brazil), respectively. By mining sequential patterns, which also involves analyzing the temporality of the data, on the records of subjects/courses taken, Yoo et al. (2017) found that students tend to drop out after taking –

usually more than once due to failure – the subjects/courses of introductory programming and mathematics. de Assis et al. (2022) performed association rules mining on student performance data augmented by social network metrics, in which the degree of propagation of Grade Point Average (GPA) was used as a proxy for the existing bond among students. As a result, the authors identified that lower school performance and minor participation in social networks lead to delayed graduation and dropout.

Using academic and demographic data from engineering students at six universities in the European Union, Prada et al. (2020) developed a web-based software tool for tutoring support. In addition to using dimensionality reduction, clustering, and visualization techniques to enable exploratory data analysis, the tool builds/uses classification models of academic performance and dropout with a descriptive objective. That is, to provide global hints about student behavior. Among the patterns found, the authors pointed out that the age of entry hurts graduation. At the same time, the admission score and the student's performance in the first semester have a positive impact on graduation.

Also using students' academic and demographic data, Yang et al. (2021) employed exploratory statistical analysis, correlation, and data mining/classification techniques to investigate how individual courses and course sequences influence student dropout/graduation on Computer Science majors at San Francisco State University. About the use of EDM, dropout prediction models based on the Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM) classifiers were developed using grade data from courses that approximately correspond to the completion of 2, 3, 4, and 5 terms, respectively. Among the results, analyzing the importance rankings of the models, the authors identified that performance in mathematics and physics courses has a more significant predictive impact at the initial stage of the degree.

Finally, Perchinunno et al. (2021) developed LR and Decision Tree (DT) models from demographic and academic data of students who completed their first year of undergraduate or single-cycle master's degrees (combination of bachelor and master) at the University of Bari Aldo Moro (Italy). The models were used descriptively to analyze the attributes and patterns related to dropouts between the first and second year at university. Among the results, the authors identified that the risk of dropping out is higher for male students or those who completed fewer than twelve credits in their first year.

In general, researchers developing predictive models are more concerned with improving results and exploring and evaluating different techniques in data pre-processing and generating classification models. Since assessing the quality of predictive models involves simulating their performance with varying evaluation metrics on a test dataset, perhaps there is greater clarity about the need for improvement and, consequently, a more effective search for techniques that help increase the performance and reliability of these models.

For example, Gamao and Gerardo (2019) and Queiroga et al. (2020) used evolutionary algorithms to optimize their classification results and evaluate different classification algorithms in developing their predictive models. In Gamao and Gerardo (2019), they proposed a swarm optimization algorithm to select an optimal subset of attributes and maximize the accuracy of predictive models. To do so, the authors considered the demographic, academic, and economic data of first-year students at Davao del Norte State College (Philippines), in addition to Naive Bayes (NB) and DT classifiers showing better predictive results. Queiroga et al. (2020) proposed the use of a genetic algorithm to optimize the hyperparameters of dropout prediction models, considering the data count of students' interactions with a Virtual Learning Environment (VLE) in a distance technical degree at Sul-rio-grandense Federal Institute (Brazil). In the proposed approach, models based on the DT, LR, RF, Multilayer Perceptron (MLP), and Adaptive Boosting (AdaBoost) algorithms, with different hyperparameter configurations, competed against each other. Thus, they used the best classifier and its best combination of hyperparameters at the end of the evolutionary process. Compared to the Grid Search optimization method and the default-configured models, the authors demonstrated better predictive results of their approach.

Barros et al. (2019) also considered technical degrees, but using academic, demographic, and economic data from face-to-face students of integrated degrees[7], at Federal Institute of Rio Grande do Norte (Brazil). In addition to optimizing the hyperparameters and evaluating the DT, MLP, and Balanced Bagging (BB) classification algorithms, different balancing techniques and evaluation metrics were tested. As a result, the BB classifier without additional balancing techniques outperformed the other models associated with varying balancing techniques. The authors also showed that it is unreliable to consider only the target class's precision, recall, and f1-score in the context of unbalanced data. It is essential to look at metrics more sensitive to performance discrepancies between classes, such as Unweighted Average Recall (UAR) and Geometric Mean (G-mean).

Using demographic and academic data from a public dataset of school students in India, Masood and Begum (2022) also evaluated different metrics and several resampling techniques for handling imbalanced data in developing LR and SVM classifiers. The authors found that SVM models constructed with random undersampling and SMOTE (Synthetic Minority Over-sampling TEchnique) in pre-processing yielded superior results. Furthermore, Masood and Begum (2022) pointed to the greater effectiveness of the AUC-ROC (Area Under the Receiver Operating Curve) for measuring predictive performance in the minority class.

As a last example, Solis et al. (2018) used academic, demographic, and economic data of undergraduate students at the Instituto Tecnológico de Costa Rica. The authors evaluated the RF, MLP, LR, and SVM classifiers, considering different perspectives on student data representation. These perspectives aimed to assess (i) whether the learning of the predictive models would benefit from including active/enrolled students among the examples of the non-dropout class and (ii) whether the dropout behavior would be better represented from the entire academic trajectory (i.e., a data record for each semester attended by the student) or from the last semester of the student. As a result, in addition to pointing out the better performance of RF, the authors identified that the best strategy was to use data from all semesters attended and only consider records of graduated/concluded students as negative examples of dropout.

Among studies that develop predictive models, those that also investigate patterns or attributes most strongly related to dropout usually do so by applying techniques that provide the importance of each feature or by analyzing the model itself when it is interpretable. Costa et al. (2021) and Naseem et al. (2022), for example, have built predictive dropout models based on data from undergraduate Computer Science students at the Federal University of Pelotas (Brazil) and the University of the South Pacific (Fiji), respectively. While Costa et al. (2021) used academic and socioeconomic data from the student's first three semesters in the development of DT, RF, and

[7] Brazilian integrated degrees combine secondary and technical education.

LR models; Naseem et al. (2022) used academic, socioeconomic, and interactional data (from interaction with a VLE) from the student's first year to build DT, RF, NB, LR, and KNN (K-Nearest Neighbor) models, considering three different prediction moments (enrollment and end of the first and second semesters). The importance rankings established by the models and the attribute selection results from the Boruta algorithm were analyzed by the first and second studies to identify the attributes with a more significant influence on dropout. In both studies, the characteristics associated with student academic performance were more important. While Costa et al. (2021) highlighted the better performance of the RF model, Naseem et al. (2022) pointed out that the NB excelled in predicting at the enrollment stage and the LR in predicting at the end of the first and second semesters.

Berka and Marek (2021) built different dropout prediction models considering academic and demographic data from face-to-face or distance bachelor's degree students at a university in the Czech Republic. The authors used data available at admission (student's first enrollment) and the end of the four initial semesters. Models were generated for each prediction moment and different attribute combinations using the DT, LR, and RF algorithms. In addition, as dropout examples, the models considered: (*i*) only students who the university dismissed for not meeting any academic rule/requirement; or (*ii*) also including students who dropped out of the degree on their own. As a result, the authors identified that the first strategy slightly improves the learning of the positive dropout class and that the LR classifiers showed better predictive performance. Furthermore, the authors interpreted the DT models to verify the influence of dropout-related attributes and patterns. They performed a dependency analysis, using different association rule algorithms and only data available at admission, to identify differences between students who did or did not follow the degree after the first semester. Among the results, the percentage of credits lost in the last semester was identified as the most essential attribute for the prediction models. At the same time, dependency analysis pointed out that the time between the end of high school and entry into higher education is the highest risk factor for early dropout.

Although less common because it is a more recent area of research, eXplainable Artificial Intelligence (XAI) techniques, particularly Shapley Additive explanations (SHAP), also appear as a resource for interpreting black box models. Baranyi et al. (2020) proposed applying a Fully Connected Deep Neural Network (FCNN) in dropout prediction. Two other deep neural network models and two ensembles of DTs were also trained and evaluated using academic and demographic data from undergraduate student admissions at the Budapest University of Technology and Economics (Hungary). The authors observed a slightly higher predictive performance of the FCNN. To better understand the decisions of their model, they evaluated the importance of attributes based on permutations and the SHAP approach. Among the results, Baranyi et al. (2020) found that the time between high school completion and higher education admission, the general admission score, and the mathematics score on the entrance exam showed greater predictive importance. Similarly, Bassetti et al. (2022) used SHAP values to measure the impact of each feature on the predictions of a Gradient Boosting Trees (GBT) model. More specifically, the authors presented ISIDE, the prototype of a student dropout alert system integrated into the online student portal of the Sapienza University of Rome. ISIDE works asynchronously, making predictions on data collected every week and returning a list with the probability of students dropping out. Before opting for the best performance of the GBT in the predictive task, the authors also evaluated DT, RF, LR, SVM, and MLP models, all built from demographic, economic, and academic data from the university's undergraduate and graduate students. Looking at the SHAP values, Bassetti et al. (2022) identified that students with more extended time since last enrollment, fewer credits earned, or lower grade averages are likelier to drop out.

4.1.2 Education Levels, Modalities and Systems

Also related to context, but answering the second research question, Figures 5(a) and 5(b) show the number of selected articles[8] by educational levels and modalities or systems, respectively. Analyzing Figure 5(a), it is possible to observe little exploration at School and Technical education levels, with the vast majority of works linked to Undergraduate. This quantity is understandable since the studies are usually developed within universities using their proprietary data. In addition, higher education institutions use academic management systems and VLEs more widely, generating greater data availability at this level of education. Regarding the modality, notice that the selected works deal mainly with face-to-face teaching[9]. However, this result may be influenced by the research interest of this review, which restricts dropout to the institutional and academic degree scopes, disregarding studies at the level of subjects/courses, which seems to be a trend in the works focused on distance modality as they explore the availability and constant updating of interactive student data in VLE. Considering Figure 5(b), it is also evident that most studies address dropout in public institutions. Most of the works presented in Section 4.1.1 follow these results, focusing on higher, public, and face-to-face education. However, it is also essential to describe the few initiatives that consider the other levels, modalities, and educational systems to exemplify and encourage the expansion of research in these contexts.

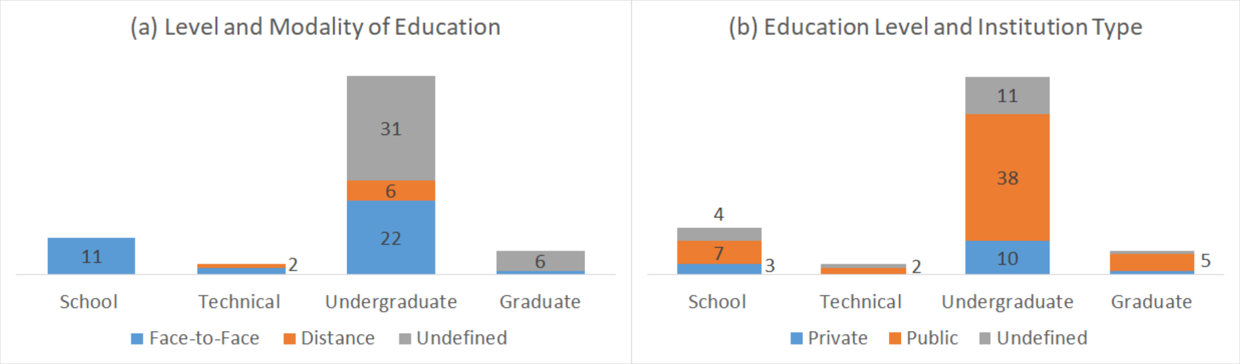


Figure 5: Number of selected articles by educational levels and modalities (a) or systems (b), respectively.

Among the eleven studies aimed at school education, Masood and Begum (2022) has already been presented. Sorensen (2019) and Orooji and Chen (2019) used academic, socioeconomic, and behavioral data obtained from public school administrative databases in the USA states of North Carolina and Louisiana, respectively. In the first study, which used data from the third through eighth grades of elementary school to predict the dropout of students enrolled in high school, SVM and boosting of DTs models showed, in that order, higher accuracy when compared to simpler prediction models. From the interpretation of DT and LR models and the feature importance

[8] Notice that the same article can be counted in more than one chart category according to the used data. This observation is valid for all graphs presented in this work.

[9] Although many works have been categorized as "Undefined" since they do not mention or give evidence of this information, these studies probably deal with the face-to-face modality, given its predominance in formal education.

of bagging DTs, the authors also identified that behavioral and academic attributes showed greater predictive power than demographic ones. Orooji and Chen (2019) evaluated different classification algorithms and balancing techniques and found that balancing techniques benefited recall.

In Chung and Lee (2019) and Lee and Chung (2019), the authors used behavioral and academic data of high school students from South Korea's National Education Information System. While in the first work, the authors balanced the data and developed an RF model, in the second one, a comparative analysis of oversampling was presented, considering the RF classifier and a boosting of DTs. As a result, in addition to validating the model's predictive performance, the authors found in the first study that attributes related to unauthorized absences and delays had the highest predictive importance, respectively. Furthermore, in the second research, the authors identified that boosting DTs without applying the balancing technique had the best results, demonstrating the need to evaluate techniques of interest before including them in a solution.

Mduma and Machuve (2021) andMduma et al. (2019b) used public datasets with academic, demographic, and economic data from students in Tanzania, Kenya, and Uganda, and only Tanzania, respectively, to develop models for predicting school dropout. Both studies employed data balancing techniques and permutation of feature importance to identify the features that contribute most to dropout. As a result, Mduma and Machuve (2021) pointed to the better performance of the LR model when compared to the MLP and RF ones, as well as the greater importance of the attributes of gender, age, income, parent check child's books, and meals per day. Although they don't present comparative or predictive performance data, Mduma et al. (2019b) signal that they have recognized an ensemble developed by soft combining tuned LR and MLP models as the best model. In addition to the attributes identified in Mduma and Machuve (2021), Mduma et al. (2019b) also pointed out as essential attributes that indicate whether the student reads books with their parent and whether their parent discusses the child's progress with the teacher. Mduma et al. (2019b) selected these essential features and considered them as input for a prototype web-based system. This system used the ensemble model and was developed to predict whether a student will drop out of school based on the information entered and to display a visualization of schools with a high risk of dropping out.

Considering the threat to the usefulness of predictive models posed by concept drifts during the COVID-19 pandemic, Xu and Wilson (2021) used imputation-based simulations to analyze the impact of data quality and availability on the performance of the dropout prediction model implemented in Rhode Island's Early Warning System (EWS). EWS is RF-based and trained on academic, demographic, and economic data of Rhode Island's public high school students, considering techniques to deal with imbalance. By building and evaluating models based on different imputation strategies and concept drift simulations, the authors found that in certain circumstances, some predictive models, while imperfect, can still be helpful in assisting decision-making.

Crespo (2020) simulated and compared the effectiveness of an income-proxy means test (PMT) and mechanisms based on dropout prediction in reaching the poor and future school dropouts for targeting social cash transfer policies. To do this, the author used academic, demographic, and economic data from Chilean governmental databases, covering students from seventh grade onwards and present in the Chilean Social Protection File. GBT, SVM, RF, elastic net, lasso, and generalized additive models were evaluated in the dropout prediction task, with elastic nets performing best. Based on the feature ranking of the models, Crespo (2020) identified greater importance in attributes related to the student's age, grades, attendance, and scholar grade/year, in addition to the previous average dropout rate in the school. The author also observed that using predictive results in conjunction with the PMT increased the targeting effectiveness, except when the social valuation of the poor and future dropouts is very different.

Queiroga et al. (2022) describe a nationwide learning analytics project in Uruguay that aims to mitigate school dropout and retention in secondary education. Using academic (covering the first grade of primary school to the second grade of secondary school) and socioeconomic data of students, the authors built eight RF models to predict dropout at different times of the first two grades (before the start of the school year and after the initial assessment meeting for each grade) of Uruguayan regular and technical basic secondary education. In addition to considering the use of balancing, attribute selection, and optimization techniques in the construction of the models, Queiroga et al. (2022) conducted a bias analysis of the models considering three protected attributes, which resulted in the approval of seven of them. The project has also developed a web API that allows the results/predictions of the approved models to be used and made available, both synchronously and asynchronously.

Like Queiroga et al. (2022), Barros et al. (2019) also considers technical and school education, taking into account the context of integrated education in the Brazilian system. Thus, in addition to being linked to the school level, they are also counted at the technical level, along with the work of Queiroga et al. (2020). Barros et al. (2019) and Queiroga et al. (2020) have already been presented in Section 4.1.1, as well as those by Perchinunno et al. (2021) and Bassetti et al. (2022), which addressed dropouts in undergraduate and graduate degrees, the same situation as Oreshin et al. (2020), Del Bonifro et al. (2020), and Fernández-García et al. (2021).

In Oreshin et al. (2020), the authors used academic, demographic, and economic data, static (available on admission) and dynamic (added semiannually), from students at the University of Information Technologies, Mechanics and Optics (Russia). In addition, interaction and sentiment analysis data from the students' publications in a social network were considered among the dynamic information. Thus, with dropout as the target variable, in addition to the early prediction model, designed to provide predictions immediately after student admission, eight prediction models were developed, considering data extracted from one to eight consecutive semesters of the degree and the GBT algorithm (chosen empirically). As a result, the authors highlighted good predictive results pointing to high school certification type, hometown, and academic center/student's area of knowledge as the top three attributes in the early model. However, for the dynamic models, the characteristics of academic performance in previous semesters showed more significant predictive potential.

Del Bonifro et al. (2020) used social and academic (school and university) data of undergraduate and single-cycle master's students from an unidentified university to build early dropout prediction models. The authors trained the models by considering only data available at admission or including data before the end of the first year. Considering random undersampling to deal with data imbalance and different combinations of hyperparameters, the authors developed models based on SVM, RF, and Linear Discriminant Analysis (LDA) algorithms for the different sets of attributes. As a result, the authors chose SVM and RF models for the scenarios with and without first-year-related features, respectively. Moreover, although both models are necessary and valuable for risk anticipation, as expected, a considerable improvement in predictive performance was observed when introducing data related to students' first year.

Given the difficulty of accessing personal data and privacy issues, Fernández-García et al. (2021) sought to develop the best possible dropout prediction models based solely on socioeconomic and academic data of undergraduate and master students from an engineering school of a Spanish public university. To do this, the authors used several feature and instance engineering techniques in the pre-processing, including balancing/resampling. Considering hyperparameters optimization, Fernández-García et al. (2021) developed GBT, RF, and SVM classifiers, along with a heterogeneous ensemble, for five different stages, specifically before enrolment and at the end of each of the first four semesters. As a differential, except for the models intended for the first stage, the others consider the prediction result of the immediately preceding stage model among their attributes. As a result, Fernández-García et al. (2021) pointed to a better performance of the GBT at the time of enrollment and of the heterogeneous ensemble in the predictions at the end of the first and second semesters.

On the other hand, Alturki et al. (2022) and Shilbayeh and Abonamah (2021) only consider data from graduate-level students, more specifically master's degrees. Alturki et al. (2022) used demographic and academic performance data of master's students in Business Informatics at the University of Mannheim (Germany). Considering SMOTE for oversampling, the authors developed LR, RF, KNN, NB, SVM, and ANN (Artificial Neural Networks) models for the tasks of predicting student status (degree completion/non-completion) and academic grade (average, above average, and below average) at the end of the first and second semesters. Among the results, Alturki et al. (2022) identified better performance of RF models trained after oversampling and that predictions made after the second semester are more accurate. In addition, using RF permutation importance, the authors identified the semester grades and the distance from the student's accommodation to the university as the most important features for predicting the student's achievements. Shilbayeh and Abonamah (2021) used socioeconomic and admission and previous academic data of master students from Abu Dhabi School of Management (United Arab Emirates). A boosted regression tree ensemble was developed to predict the number of enrolled students in subsequent academic years. In addition, the authors applied the Apriori algorithm to extract association rules and discovery patterns about master's students and who is most likely to drop out. As a result, Shilbayeh and Abonamah (2021) highlighted the superior performance of the ensemble when compared to a single boosted regression tree, and characteristics related to the 30-40 age range and the 2.5-3 undergraduate GPA range showed recurrence in the rules associated with dropping out.

Queiroga et al. (2020), Berka and Marek (2021), Kang and Wang (2018), Yu et al. (2021), and Ortigosa et al. (2019) are examples of articles that consider the context of distance learning. Kang and Wang (2018) and Yu et al. (2021) used academic and demographic data of undergraduate students at USA public universities to develop models for predicting dropouts during winter break and the first year of the degree, respectively. Kang and Wang (2018) developed logistic prediction models using or not using data random undersampling. As each model favored accuracy or recall, authors used both to predict the dropout risk of active students, generating two listings of at-risk students for educational managers. In Yu et al. (2021), the authors produced separate models for students in distance and face-to-face degrees based on the data available at the end of the first period/term of the degree. The models were generated from the LR and GBT algorithms, considering or not the inclusion of sensitive attributes, such as gender and color/race, to assess whether these data produce discriminatory models. As a result, after evaluating the overall performance and fairness of the predictions (variation of predictive performance on minority groups) of each model, the authors identified that the GBT models performed better and that sensitive attributes had little impact on the algorithmic fairness of the predictions, being preferable to use them.

In addition to describing the development and evaluation of prediction models, Ortigosa et al. (2019) presented the challenges and technical changes imposed by implementing an online dropout risk prediction system for undergraduate students at Universidad a Distancia de Madrid (UDIMA). The system provides (*i*) static early predictions, which consider data available on student enrollments in their respective academic years/cycles, and (*ii*) dynamic predictions, which include data from student interaction in the VLE, collected and updated periodically. These predictions are based on demographic, economic, academic, and interactional data. By dividing the academic year into ten monthly periods, the authors developed 22 specialized models (one static and ten dynamic for each type of student – new or recurrent) based on the C5.0 DT algorithm. In addition, the authors identified that age and university access type were considered necessary attributes in the prediction models of new students but ignored by those of veterans, who began to consider attributes related to student performance in previous years/cycles. Following this behavior, the greater the data the models consider, the better the results presented.

As well as Ortigosa et al. (2019), another 13 works consider data from the private system. Perez et al. (2018) and Hannaford et al. (2021), for example, devoted their research to specific undergraduate degrees. They developed predictive models for the dropout of Systems Engineering students and the risk of not graduating Nursing students, considering private universities in Colombia and the USA, respectively. Perez et al. (2018) developed DT, LR, NB, and RF models utilizing academic and demographic data of the students. Hannaford et al. (2021), in turn, developed 126 predictive models, considering different combinations between demographic and academic data (from high school and university). Models were built for five prediction moments, considering the beginning of additional academic years, based on eight classifiers and one ensemble based on the weighted average of the results of these classifiers. As a result, Perez et al. (2018) identified that the RF model showed superior performance, while the ensemble with eight classifiers scored better in Hannaford et al. (2021). The latter work also pointed out that the more advanced the prediction moment, i.e., the more academic data is considered, the higher the model's performance. In common, both studies identified that data related to university academic performance have greater predictive importance. The most essential attributes included cumulative GPA and GPA in the degree-specific subjects/courses.

Hutagaol and Suharjito (2019) also proposed to combine the results of three prediction models (KNN, NB, and DT) into an ensemble; however, using a GBT meta-classifier to predict the dropout of students from the Faculty of Social and Political Sciences of a private Indonesian university. Starting from demographic, economic, and academic data, the authors performed data balancing and attribute selection in pre-processing by considering SMOTE and Learning Vector Quantization (LVQ) techniques, respectively. As a result, in addition to the ensemble outperforming the individual classifiers, LVQ highlighted the greater predictive importance of attributes related to students' attendance and academic performance.

da Silva et al. (2019) used demographic, academic, and economic data of students from onsite degrees at public and private universities provided by the Brazilian Census and Higher Education Flow Indicators. The authors built and evaluated three ensembles of homogeneous regressors, each corresponding to the bagging of linear, robust, or ridge regressors, in addition to reproducing a heterogeneous ensemble proposed by another work, with seven base algorithms combined by a Ridge meta-regression. The authors trained these four regression ensembles from two distinct subsets of attributes, selected by Stepwise and Pearson correlation methods, respectively. As a result, the homogeneous ensembles outperformed the heterogeneous ones, with emphasis on the bagging of linear regressors trained on the attributes selected by the Stepwise method, which had lower error rates. In addition, the degree retention and completion rates, the number of students remaining, and the study shift were identified as the most significant predictive importance for dropout models.

In Palacios et al. (2021), the authors developed models to predict dropout at the global level (i.e., independent of the period) and specialized in the first, second, or third year of the degree. Using academic (school and university), demographic, and economic data of undergraduate students at Universidad Católica del Maule (Chile), in addition to the SMOTE balancing technique, the authors trained DT, KNN, LR, NB, RF, and SVM classifiers for the four prediction objectives. However, only the models designed to predict dropouts in the second and third years considered student performance data at the university. As a result, the authors highlighted the importance of data balancing and good predictive performances, especially of RF models. Moreover, through information gain analysis, the authors identified that the attributes related to the student's school performance and the socioeconomic background of the place of origin are the most important for the first three scenarios. However, university performance attributes also proved relevant in the third one. On the other hand, attributes related to students' university performance proved to be more critical for predicting dropping out in the third year of the degree (fourth scenario).

As a last example, Shiau (2020) addressed the dropout problem of full-time undergraduate students on Taiwan's private and technological university campus. The author used a decision table and a DT model in the descriptive analysis of dropout determinant rules to help establish counseling strategies and identify at-risk students. Both techniques related the dropout rules, for the most part, to students' low academic performance. Considering the DT model, the author identified, for example, that first and second-year students with no tuition discounts and average academic performance lower than 45.6 in the last semester tend to drop out. In addition, to provide early predictions, Shiau (2020) developed a model based on the discriminant analysis method and data on student absences. The resulting function was shown to be most influenced by absences related to sick leave. Therefore, it has been incorporated into a form allowing academic advisors to enter students' up-to-date absence figures and get their dropout risk in real time.

4.2 Data

Regarding the perspective of the data used and the third research question of this RSL, Figure 6(a) shows the distribution of the selected articles according to the types (nature) of attributes and educational levels that they considered. It is necessary to mention that information related to student absences was classified as academic in this categorization, although some works cited in Section 4.1 refer to this data as behavioral. In addition, some studies used complementary attributes extracted from questionnaires/surveys, such as daily time for leisure and homework, which were not considered in the categorization. This fact is because, in addition to not being a common practice among the works, data resulting from questionnaires/surveys are subject to limitations and biases, as mentioned in Section 1.

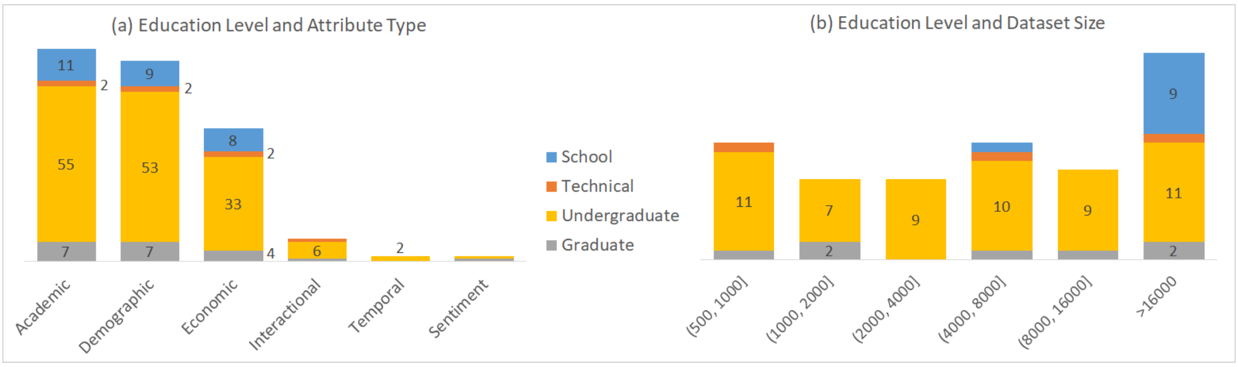


Figure 6: Number of articles by attribute type (a) and dataset size (b).

Considering that we hide the labels related to a single article so as not to pollute the graphs, notice that the Undergraduate level is the only one to have works linked to all data types, which is understandable since it is the most researched level. Academic, social, and economic data are widely used in predicting institutional and educational degree dropout, as opposed to interactional, temporal, and sentiment analysis data. Although not shown in Figure 6(a), it is worth specifying some attributes pointed out by the studies as being of greater predictive importance. Among the academic qualities, we can highlight those related to academic performance, such as admission exam grades, and grade point and frequency averages (in the term or accumulated), and to the student's position in the degree, such as the number of terms/semesters taken, and the total number of credits earned or lost (in the term or accumulated). Regarding demographic data, the attributes of age at entry, gender, and time between finishing high school and entering higher education stand out, in addition to those related to the students' origin, such as hometown and indicators of socioeconomic background. Finally, among the economic information, we can mention family income attributes and student aid or benefits receipt.

While Naseem et al. (2022), Oreshin et al. (2020), Ortigosa et al. (2019), Queiroga et al. (2020), Deho et al. (2022), Park and Yoo (2021), and Villegas-Ch et al. (2020) correspond to articles that used the interactional data type, the work of Oreshin et al. (2020) was the only one to use predictive attributes resulting from sentiment analysis. However, it is related to both undergraduate and graduate levels. Yoo et al. (2017) and Kurniawati and Maulidevi (2022), in turn, considered in their analysis the temporality associated with students' academic records[10].

Using academic, demographic, and economic data from undergraduate students at an Indonesian university and taking advantage of the semester temporality/sequentiality of the educational path data, Kurniawati and Maulidevi (2022) developed recurrent neural network models, more specifically Long-Short Term Memory (LSTM) and Gate Recurrent Units (GRU) classifiers, to predict graduation/dropout and final GPA (below average, average, good, very good). As a differential, the authors built three models: one model for each prediction task (dropout and final GPA), constituting a separate method; and one model to predict the data with dropout and final GPA cross-labels, constituting a combined method. Evaluating the predictive performance of the models over different semesters, Kurniawati and Maulidevi (2022) found that GRU and LSTM

[10] Other studies have used temporal data, such as the semester in which the student is enrolled or the interval between the end of high school and entry into higher education. Still, they have not extracted information based on the sequentiality of these data, nor have they considered their time series.

outperformed in graduation/dropout and GPA prediction, respectively. In addition, they pointed out that the models showed better results in the separate method, with satisfactory predictive performance from the first and second semesters for the graduation and GPA models, respectively.

Also considering data characteristics, Figure 6(b) shows the relationship between the selected articles and the sizes of the datasets (number of students/records) used, considering each education level separately. Although the works are based on datasets of quite varied dimensions, it is clear that studies related to the school education level mostly concentrate on the broadest range of data. This fact is understandable as these studies usually use data from all schools that make up the public system of a state or country, as seen in Section 4.1.2. Similarly, the other works in this range, such as da Silva et al. (2019), Yu et al. (2021) and Oreshin et al. (2020)[11], generally using undergraduate student data from all degrees or several departments of an institution.

As exceptions, Beaulac and Rosenthal (2019) and Viloria et al. (2019) considered only data from the Faculty of Arts and Science at the University of Toronto and the Departments of Engineering at the University of Mumbai, respectively. In more detail, Beaulac and Rosenthal (2019) used first-year academic performance data to build models from the RF and LR (baseline) algorithms to predict whether students would get their degrees. As a result, the authors confirmed the better performance of the RF model, pointing out the predictive importance of attributes associated with credits and grades in a seminar subject/course and other subjects/courses related to specific departments, especially mathematics. Viloria et al. (2019), in turn, used academic, demographic, and economic data to build predictive models of dropout based on Bayesian Networks (BN), ANN, and DT algorithms. In addition to considering the results of all models as satisfactory, the authors confirmed, through the interpretation of the models or their attribute rankings, the hypothesis that academic and socioeconomic conditions (such as scores on the entrance exam and student benefits, respectively) are decisive for the student permanence or dropout.

On the other hand, works that consider datasets of more restricted size (range of 500 to 1000 students/records) generally address dropout prediction in a single degree, such as Costa et al. (2021), Hannaford et al. (2021), Naseem et al. (2022), Perez et al. (2018), Queiroga et al. (2020), and Yoo et al. (2017). However, there are exceptions, such as the Iam-On and Boongoen, 2017b (2017a, 2017b) studies, in which the datasets, although small, cover academic and demographic data of students from 26 academic departments at Mae Fah Luang University—as in the first work, described in Section 4.1, Iam-On and Boongoen (2017b) considered separately the contexts of students who had just entered and completed the first year of the degree. However, in this second research, the authors developed predictive models of dropout to validate a proposed data transformation approach based on the Link-Based Cluster Ensemble. As a result, despite requiring more computational effort, the authors demonstrated that their approach improved predictions and outperformed other dimensionality reduction techniques, considering models developed for both contexts from the DT (C4.5), KNN, NB, and ANN algorithms.

The relationship between the articles and the nature/type of the attribute used, now considering the perspective of the education modality, is presented in Figure 7(a). Notice that adopting data types for face-to-face teaching generally follows the same trend as for all selected studies. However, interactional data show the same representativeness as economic data in distance learning, being considered by 57% (4 out of 7) of the works in this modality, which confirms the greater

[11] Notice that the graph counts the Oreshin et al. (2020) study at both undergraduate and graduate levels.

exploration of student interaction with VLEs in this context. Additionally, in Figure 7(b), we can see the mapping about the amount of data types considered in the articles. Notice that most studies use three types of data, followed by those adopting two or one type.

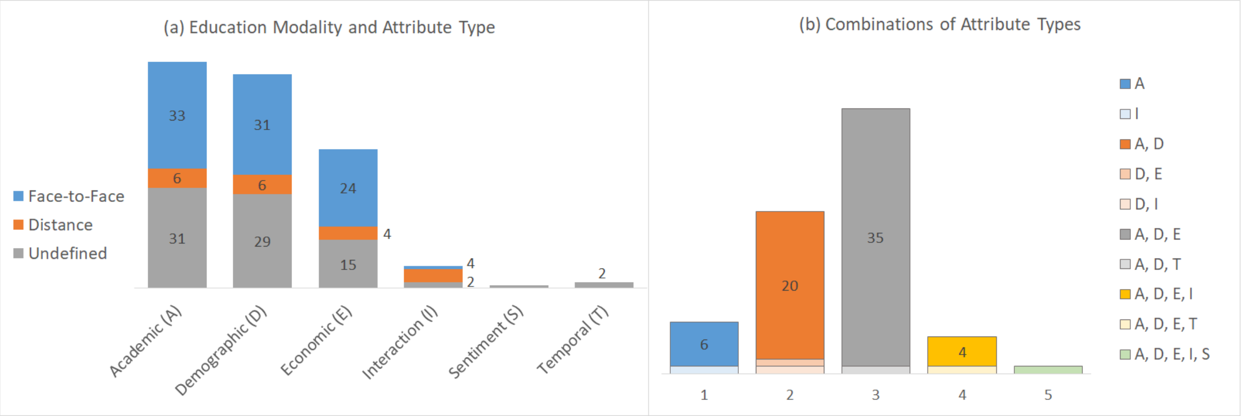


Figure 7: Total articles by data types and modalities (a); and by combinations of types (b).

We identified the use of four or five types of data only in Ortigosa et al. (2019), Naseem et al. (2022), Park and Yoo (2021), Villegas-Ch et al. (2020), Kurniawati and Maulidevi (2022), and Oreshin et al. (2020). Most of these studies consider academic, demographic/social, economic, and interactional information (A, D, E, I), with the exception of Kurniawati and Maulidevi (2022), which considers temporal rather than interactional data (A, D, E, T), and Oreshin et al. (2020), which adds sentiment analysis data to the first four types (A, D, E, I, S). The works linked to three types generally used academic, demographic/social, and economic data (A, D, E), as in Barros et al. (2019), Costa et al. (2021), Gamao and Gerardo (2019), and Solis et al. (2018). Yoo et al. (2017) was an exception and used academic, demographic, and temporal (A, D, T).

In works related to two types, as in Berka and Marek (2021), Iam-On and Boongoen, 2017b (2017a, 2017b), and Perchinunno et al. (2021), the authors generally used a combination of academic and demographic data (A, D). The exceptions correspond to Deho et al. (2022) and Freitas et al. (2020), who adopted demographic and interactional (D, I) or economic (D, E) data, respectively. It is worth mentioning that Freitas et al. (2020) developed an IoT (Internet of Things) system for predicting early dropout of engineering students at Instituto Federal do Ceará, Campus Fortaleza (Brazil). Using only seven socioeconomic attributes, the authors built LR, DT, SVM, KNN, DNN, and MLP models. An interface has been provided to users so that they can, among other facilities, select the model to be used in the prediction, submit student records to be predicted or consult the risk of students with stored data. Regarding the predictive results, the authors pointed out better performances for the DT and DNN models.

Finally, when the works used only one type, the academic (A) stood out, as in Beaulac and Rosenthal (2019), Chung and Lee (2019), Lee and Chung (2019), Rovira et al. (2017), de Assis et al. (2022), and Kuzilek et al. (2021). The only exception was the Queiroga et al. (2020) study, which considered just interactional data (I), as described in Section 4.1.1.

4.3 Tools and Techniques

Answering the first item of the third research question, the articles selected in this RSL addressed dropout prediction mainly through the Classification task, either to identify at-risk students or to determine attributes and patterns related to dropout. The exceptions were the Iam-On and Boongoen (2017a), Yoo et al. (2017), da Silva et al. (2019), de Assis et al. (2022), and Shilbayeh and Abonamah (2021) studies, which used only Clustering, Association, or Regression techniques.

We categorized the algorithms used in the works according to the methods/techniques they represent. Figure 8(a) shows the frequencies of the categories that showed recurrence. We can see that the authors adopted ensemble methods in most studies, specifically in 48. Ensemble algorithms have gained attention in recent years because they combine multiple classifiers and usually increase the quality of predictions due to the greater generalizability of this set of models (Lee & Chung, 2019). Although not shown in the graph, we should mention that most occurrences of the ensembles (43) considered DTs as base classifiers, with emphasis on the high frequency of the RF algorithm used, for example, in Chung and Lee (2019), Solis et al. (2018), and Queiroga et al. (2022). Also, we can see that DTs are not only applied as base classifiers. Their applications, as in Freitas et al. (2020), Perez et al. (2018), and Ortigosa et al. (2019), correspond to the second most recurrent category of an algorithm. The reason may be associated not only with their good predictive and computational performance but also with the interpretability of their results. This is because the DTs are constituted by rules describing the discovered patterns (Han et al., 2012). A similar motivation may account for the fact that linear models, used in Kang and Wang (2018), Mduma and Machuve (2021), Perchinunno et al. (2021), and Shiau (2020), for example, appear as the third most frequent category, with emphasis on LR models.

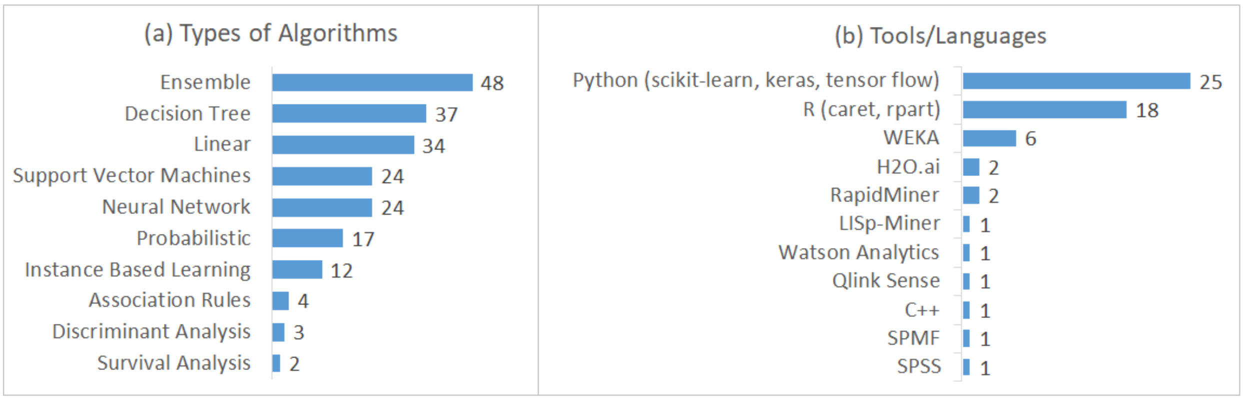


Figure 8: Articles by categories of algorithms (a) and tools/languages (b) used.

Still, regarding Figure 8(a), it is evident that categories related to the SVM, neural network (simple or deep), and probabilistic (based on Bayes' theorem) algorithms also had significant frequencies. The category of instance-based learning, represented mainly by the KNN algorithm, had a more discrete recurrence. Freitas et al. (2020), Hannaford et al. (2021), and Palacios et al. (2021) are examples of works that used KNN, among other algorithms, in their comparative analyses. Although it is also a traditional algorithm, this lower adoption of KNN may be associated with the computational cost it requires in classifying new instances, which in the worst case may increase linearly with the size of the training dataset (Raschka & Mirjalili, 2017). Finally, the other categories were slightly used, having been cited in less than five studies.

G. Santos et al. (2020) considered academic and demographic data of undergraduate students from a Brazilian federal institution in developing a dropout prediction model named EvolveDTree.To improve the learning of the DT classifier in the face of the unbalanced dataset, the authors employed the K-Means clustering algorithm in pre-processing. Thus, the data was stratified into ten groups of records, which were used in training by cross-validation. Furthermore, to avoid overfitting the DT model, an attribute selection method based on an evolutionary (genetic) algorithm was also used, resulting in eight selected attributes: an indicator of participation in social programs, degree closure semester, writing grade in the entrance exam, year, and semester of admission, GPA, race/ethnicity, and the number of years in the degree. In addition to validating the results of each step of their approach's results, the authors demonstrated their model's superior performance over KNN, AdaBoost, MLP, RF, NB, SVM, and Quadratic Discriminant Analysis (QDA) models. Finally, the visualization of the model also showed that EvolveDTree considered potential dropouts, mainly students with lower academic performance or who had not yet passed the second year of the degree.

Chen et al. (2018) proposed using survival analysis, based on Aalen's Additive and Cox's Proportional Hazard models, to predict the occurrence and timing of undergraduate student dropout. The research used demographic and academic (school and university) data of students at George Mason University (USA) who were linked to Science, Technology, Engineering, and Mathematics (STEM) fields. The models were trained incrementally, considering the information available in different periods, from the data available at the end of the second semester to the data available at the end of the fifth semester. The authors compared their survival analysis models to traditional machine learning models (LR, DT, RF, NB, and AdaBoost). As a result, in addition to observing that the predictive results improved as new semester data were considered, they identified that survival analysis models performed better when a few semester data were available. Additionally, the attributes of semester-wise GPA and the number of semesters attended/enrolled showed greater predictive capacity.

Figure 8(b) shows the tools and programming languages used by the selected articles in the data mining process. Although some works do not mention the tools adopted, analyzing the graph makes it possible to see the wide use of Python and R languages. It is necessary to say that the use of Python appears in general associated with the machine learning library Scikit-Learn, such as Barros et al. (2019), Queiroga et al. (2020), and Rovira et al. (2017). In the Rovira et al. (2017) work, models were developed to predict, among other targets, the risk of undergraduate students at the University of Barcelona not re-enrolling in the second and third year of their degree. The authors used the academic performance data of students in the first year of Law, Computer Science, and Mathematics degrees separately. Five prediction models were developed for each degree, considering the LR, SVM, RF, NB, AdaBoost classifiers, as well as the SMOTE balancing technique. As a result, the RF and AdaBoost models performed best in the Law degree, while the LR and NB models excelled in the other two. Considering the sizes of the datasets used for each degree, the authors pointed out that non-parametric classifiers may be better suited to the context of low data availability. In contrast, parametric models may be better adjusted to the opposite situation.

Similarly, the Caret package is frequently used by studies adopting the R language, such as Perchinunno et al. (2021), Lee and Chung (2019), and Solis et al. (2018). Among the other tools, only WEKA, RapidMiner, and H2O.ai were referenced more than once by the selected studies. Palacios et al. (2021), Viloria et al. (2019), and Vega et al. (2022) are examples of research using WEKA. The only two references to RapidMiner occurred in Berka and Marek (2021) and Nagy and Molontay (2018). Although Berka and Marek (2021) used RapidMiner in the classification, the authors also applied the LISp-Miner tool in the association mining. Similarly, Nagy and Molontay (2018) has also used implementations of the H2O.ai platform, also adopted by Deho et al. (2022), in some of its models.

Based on academic and socioeconomic data of undergraduate students from the Faculty of Systems Engineering and Informatics of the Universidad Nacional Mayor de San Marcos (Peru), Vega et al. (2022) developed a DT model and a (not web) system for dropout prediction. The authors used six feature selection methods available in the WEKA tool to select the features frequently considered necessary, and applied SMOTE to deal with the data unbalanced problem. As a result, Vega et al. (2022) presented a diagram of use cases and interface prototypes related to the system developed, highlighting the importance of the cumulative and last cycle weighted average grades and the number of credits earned features.

Nagy and Molontay (2018) developed early dropout prediction models from academic and demographic data available when undergraduate students were admitted at the Budapest University of Technology and Economics. In addition to using different attribute selection strategies (evolutionary, by correlation, and based on feature importance metrics dependent on GBT and DNN models), the authors considered several classifiers, specifically those of linear regression, DT, RF, GBT, LR, NB, KNN, DNN, AdaBoost. Among the attributes selected as important, the authors identified the performance in humanities subjects/courses in high school. This selection suggests that a good and broad education is essential for university success, even for engineering and exact sciences students. Regarding the models, GBT and DNN performed better on the selected and complete sets of attributes, respectively. As a final result, Nagy and Molontay (2018) developed a web application to provide predictions from the GBT model, trained from the reduced/selected attribute set. Thus, by entering the attribute data of a specific student, the user can obtain the prediction of their risk of dropping out right at the time of enrollment.

In addition to the technical characteristics already presented, it is crucial to position the studies concerning some additional aspects. For example, Figures 9(a) and 9(b) show the proportions of works concerning the adoption of balancing and attribute selection techniques, respectively. These analyses are necessary because unbalanced data tends to be a reality in the context of dropouts, and this inequality can affect the learning of the underrepresented class of students. In addition, attribute selection techniques aim to remove irrelevant or redundant dimensions from the data, reducing the processing cost and preventing these variables from negatively influencing the models Han et al. (2012).

Notice that only 41% (29) of the works cited the application or evaluation of balancing techniques. Although some databases are naturally balanced, this low percentage shows that many studies have neglected this critical issue. Among the works that balanced their training data, the use of the SMOTE technique stands out, as in Hutagaol and Suharjito (2019), Lee and Chung (2019), and Rovira et al. (2017). Attribute selection techniques were considered in 35% (25) of the studies. This result is understandable since many works already start from a reduced number of attributes, either due to a lack of access to more information or because they base the choice of attributes on previous experiences or research. Among the selection approaches used in the works already described, we can mention: evolutionary methods, such as Gamao and Ger-

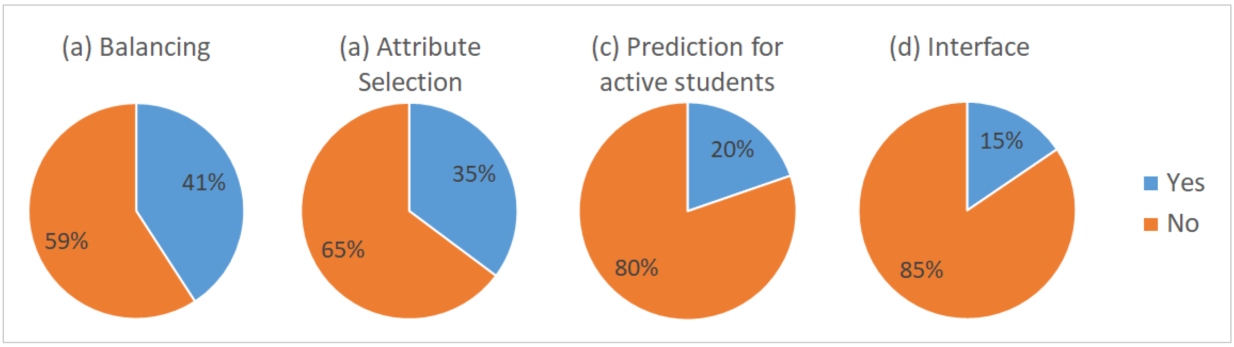


Figure 9: Proportions of articles concerning technical characteristics.

ardo (2019) and G. Santos et al., 2020; by correlation and stepwise, such as Fernández-García et al., 2021 and da Silva et al., 2019; LVQ in Hutagaol and Suharjito (2019); selection based on model-dependent importance metrics, in Nagy and Molontay (2018); and Boruta algorithm, which operates as an RF wrapper, in Naseem et al. (2022).

Another example of a study applying attribute selection is Tsai et al. (2020). Using demographic, economic, and academic performance data from first-year students at a Taiwanese university, the authors developed models to predict dropout over the next three years. The authors performed a statistical analysis based on logistic regression to attribute selection. This analysis pointed to four main attributes: academic performance (class ranking percentage), student loan application indicator, number of absences, and number of subjects/courses in which the student received academic performance alerts. Considering these four attributes for each student's first two semesters, MLP and LR classifiers were trained and showed better specificity and sensitivity, respectively. As a final result, the authors used the most sensitive model in the prediction task, providing the university with lists of at-risk students, thus enabling the execution of preventive strategies.

Figure 9(c) shows the proportion of papers that automatically predicted the risk of dropout for enrolled (active) students. Although many studies (80%) applied and validated prediction models on historical data to identify patterns related to dropout or define the best prediction algorithm, only 20% (14) of the works applied the models built on current data to predict and detect future dropout automatically. Specifically, Kang and Wang (2018), Nagy and Molontay (2018), Ortigosa et al. (2019), Mduma et al. (2019b), Freitas et al. (2020), Shiau (2020), Tsai et al. (2020), Shilbayeh and Abonamah (2021), Xu and Wilson (2021), Bassetti et al. (2022), Queiroga et al. (2022), Vega et al. (2022), Demeter et al. (2022), and Urbina-Najera and Mendez-Ortega (2022) correspond to the latter case.

Interested in the predictive utility of financial aid records, Demeter et al. (2022) used academic, demographic, and economic data from undergraduate students who applied for free federal aid in their first two consecutive years of enrollment at a public university in the southeastern USA. In addition to selecting attributes from a theoretically oriented process, the authors used supporting analytical and data-driven techniques, such as stepwise logistic regression, and the model's feature importance. As a differential, Demeter et al. (2022) used a two-level hierarchical classification process to make predictions based on records available at the end of the second to sixth semester after matriculation. This process comprises two RF models, one to initially predict whether students would drop out or graduate, and another to be applied to expected graduates to predict whether they would complete their degree late or on time (4 years). In addition to validating the predictive performance of their models, the authors applied them to actual student data. They identified a list of potential dropouts that was shared with academic advisors for their analysis. Demeter et al. (2022) also highlighted the predictive importance of attributes related to credit earned, college and high school GPA, estimated family financial contribution, and enrollment and grades in required gateway courses within a student's major.

Urbina-Najera and Mendez-Ortega (2022) used academic, demographic, and economic data from undergraduate students in 25 engineering, social science, and administrative science programs at an unidentified university. By oversampling the minority class (dropouts) and undersampling the majority class (non-dropouts), the authors evaluated different feature selection methods (RF wrapper, correlation, and consistency-based) in the development of RF and ANN models for dropout prediction. Among the results, they pointed out that resampling benefited the predictive ability of the models and that the RF with features selected by the RF wrapper showed better performance. Urbina-Najera and Mendez-Ortega (2022) also reports that the (best) RF model has been implemented in the institutional system and has allowed predicting an approximate number of possible dropouts per period, contributing to the preventing actions.

Another factor to highlight is that not all the studies that predict future dropout made their predictions available to academic managers through interfaces, as indicated in Figure 9(d). That is, although 92% of the works developed predictive models for dropout (Figure 4), only 15% ensured intuitive and continuous access to the actual predictions, facilitating monitoring and development of preventive strategies for at-risk students. In addition, it is essential to mention that among the 11 works that provided systems/interfaces, only the ones by Ortigosa et al. (2019), Xu and Wilson (2021), and Bassetti et al. (2022) involve projects that integrate the prediction systems with the databases of administrative/institutional systems. The others require the user to inform the students' data by filling out a form or submitting a .csv file to have their status predicted.

5 Trends, Opportunities, and Challenges

The analysis of the studies covered by this SLR allows us to observe some trends, opportunities, and challenges related to the application of EDM and machine learning in predicting student dropout, considering the degree and institutional levels.

From the contextual perspective, we can highlight the lack and consequent opportunity for research at school (primary and secondary) and technical education levels and in countries with the lowest HDI, which are precisely the contexts with the greatest need. This is because basic education levels cover the most significant number of students and are extremely important for the socio-economic and educational development of the population. Furthermore, it is necessary to consider the challenges that may be at the root of this lack and that need to be addressed to drive progress in this research context. Although they concentrate most of the students, public schools have limited resources. They may not have dedicated teams specializing in information technology, making it difficult to conduct this type of research/study locally (within the schools). Therefore, expanding the use of EDM and machine learning in the school dropout context depends on government bodies investing in the development of this research internally or facilitating it externally by making public datasets available, preferably covering data on different aspects of students[12]. Mduma and Machuve (2021) also pointed out the difficulty of finding publicly available datasets that address the problem of student dropout for the development of studies involving the application of machine learning techniques. Also in line with the remarks on the school context, in this SLR, two of the three studies related to technical education were carried out in Brazil, using data from institutions of the Federal Network of Professional and Technological Education, which also have undergraduate and graduate degrees, providing specialized technical staff and a more conducive environment for research.

Regarding the data used, we observed that many studies utilize a combination of demographic, economic, and academic data despite the widespread use of VLEs in universities. This shows a research opportunity, as analyzing student interaction data in VLEs can aid in identifying behavioral patterns. Furthermore, they can be precious in early prediction models since they are accessible throughout the entire semester, even before the student has completed any academic terms and obtained GPA information (Park & Yoo, 2021). Another opportunity behind the extraction and use of interactional data is that the texts of students' posts and their participation in forums can be processed and serve as input for a sentiment mining process, the result of which can also contribute to the prediction of dropouts. Similarly, further exploration of the temporality and sequentiality of academic trajectory data is needed. Studies comparing the predictive results of traditional classification models with those of time series would be of great practical use. They would shed light on whether the temporality of the data truly offers a predictive advantage in the dropout context.

In general, the studies on dropout prediction also show little adaptation of techniques to the specific domain. Although the research questions are focused on the dropout problem, the studies generally use traditional data mining and machine learning techniques without adapting them or their applications to the domain's particularities. Fernández-García et al. (2021) and Demeter et al. (2022) are examples of studies that propose different classification structures and can serve as inspiration for other research in this direction. It is also common to develop specialized models for degrees or prediction periods. Although exploring more specific factors, such as student performance in particular subjects/courses, may benefit such models, it is crucial to notice that this approach may provide less comprehensive results institutionally. Therefore, when possible, developing models that focus on the generic prediction of dropout (i.e., prediction models intended to predict dropouts in different degrees and periods/grades of the academic trajectory) may have greater returns for educational practice.

Still considering the technical perspective, although we observed the concern to evaluate different machine learning algorithms (this is indeed important) for the development of new studies, we should highlight the importance of validating the technical decisions throughout the entire model development process, considering the institutional reality and the objectives of each research. In developing a generic prediction model, for example, is it best to consider absolute attributes (such as the number of completed credits)? Would not using relative attributes (such as the percentage of completed credits) provide more informative and generalizable data for different curricular structures? Studying the choice and prioritization of evaluation metrics is also extremely necessary. Otherwise, all technical decisions may be based on biased results, generat-

[12] This is because much public data, such as that from educational censuses, is limited to socioeconomic information on students, which limits the research potential and, consequently, the interest of researchers.

ing falsely viable models. Although many studies point to the good performance of their models, these conclusions are sometimes not adequately substantiated. For example, classification models should not be evaluated solely based on their overall hit rates (accuracies), especially if data imbalance has not been deal with. In addition, although DT-based models have been widely used because they combine good prediction and execution performance with interpretability, it is worth noting that there are several possibilities for applying more complex models, such as deep neural networks. This is because if these models provide a more significant predictive advantage and acceptable execution times for the reality and needs of the project/research, XAI techniques can be used to identify the factors contributing to the prediction of dropout (Rondado de Sousa et al., 2021).

In addition, while there is growing interest in using EDM to predict dropout, most studies only analyze historical data. They don't often apply the models to active/enrolled student data to predict who is at risk of dropping out. Furthermore, not all of the few works that perform predictions provide academic managers with interfaces/tools for intuitive and continuous access to the results of the predictions. This would be important to support decision-making and planning of preventive strategies. We believe that this underutilization of the efforts and potential of many studies in educational practice is the greatest challenge. This is because, as already mentioned, most of the research is carried out in an academic environment, often associated with degree final works, dissertations, or theses. In this type of research, due to privacy issues, access to data is usually completely anonymous and partial, with no continuous access to databases. As a result, these studies generally do not allow for adequate progress in identifying at-risk students. In our view, the solution to this problem depends on carrying out projects involving researchers and stakeholders linked to the information technology teams and administrative managers of the institutions that hold the data. In this way, models can be developed in the academic sphere and integrated into institutional systems with the effort of the partner institution.

Finally, we must point out the limitations and threats to the validity of this RSL. First, due to the large number of papers collected, we did not use peer review to select and analyze the articles. In other words, each publication was examined by a single researcher. In the interpretive context inherent in any secondary study, this can lead to subjective decisions. Another threat could be the lack of scientific databases that are important for the specific contexts of different countries. However, since this work aims to investigate the international scenario and we don't know the relevant databases for each country, including some specific databases could introduce bias in the results. For this reason, we decided to consider only four major international scientific databases, namely ACM Digital Library, IEEE Xplore, Web of Science, and Scopus.

6 Conclusion and Final Remarks

This work presents a systematic review of the application of EDM in the context of student dropout prediction. This review considered the scopes of institutional and academic degree dropout and identified contextual, technical, and data characteristics addressed in this research topic.

The selected articles focused more on the undergraduate level, the public education system, and the face-to-face modality. This last trend may be related to the scope of this work, which disregards publications associated with the prediction of dropout in subjects/courses that are recurrent in the distance learning modality. Regarding the information used, besides a remarkable variation in the number of records considered, almost all studies adopt from one to three types of data in the representation of students. Combinations of academic, social, and economic attributes are the most frequently used. Although socioeconomic attributes have been highlighted as necessary by many studies, mainly for the early prediction of dropouts, several studies have pointed out that, when available, data on students' academic performance overlap other types of attributes, dominating the decisions of predictive models.

From the perspective of the techniques and tools used, most works address the problem of dropout prediction through the classification task, using DT models, individually or in an ensemble. The use of Python and R languages also stands out in conducting the EDM process. In addition, although we already expected limited adoption of attribute selection techniques, considering that many studies start from a small number of variables, the low application of balancing techniques caused a surprise. This is because dropout datasets usually have a significant imbalance between classes, which can impair the quality of the models.

Still, on the technical characteristics, although interest in adopting EDM in dropout prediction is growing, most works only analyze past data. They do not apply the models to data of active/enrolled students to predict who is at risk of dropping out. Moreover, not all of the few works that perform predictions provide academic managers with interfaces/tools for intuitive and continuous access to the predictive results. This is important to assist in decision-making and planning preventive strategies. So, these results indicate the underutilization of the efforts and potential of many studies in the educational practice.

Finally, at the end of this study, we also presented and discussed research opportunities and challenges that follow the general trends observed among the works analyzed. Despite the limitation of not conducting a peer review in carrying out the SLR, we hope that the results and findings of this work will contribute to the development of new studies related to the application of the EDM in predicting student dropout. In addition to instigating the expansion of this area of research in underdeveloped countries and greater caution regarding technical decisions, we hope to encourage the more practical application of predictive models to enrolled student data and the availability of these predictions to academic managers. This will undoubtedly add to the quality and usefulness of the work in educational practice. These observations will also guide the continuity of our research.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001. The authors thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Universidade Federal de Pelotas, and Instituto Federal de Educação, Ciência e Tecnologia Farroupilha.

References

Agrusti, F., Bonavolontà, G., & Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of e-Learning and Knowledge Society*, *15*(3), 161–182. https://doi.org/10.20368/1971-8829/1135017 [GS Search].

Agrusti, F., Mezzini, M., & Bonavolonta, G. (2020). Deep learning approach for predicting university dropout: a case study at Roma Tre University. *Journal of e-Learning and Knowledge Society*, *16*(1, SI), 44–54. https://doi.org/10.20368/1971-8829/1135192 [GS Search].

Aguirre, C. E., & Pérez, J. C. (2020). Predictive data analysis techniques applied to dropping out of university studies. *2020 XLVI Latin American Computing Conference (CLEI)*, 512–521. https://doi.org/10.1109/CLEI52000.2020.00066 [GS Search].

Alturki, S., Cohausz, L., & Stuckenschmidt, H. (2022). Predicting master's students' academic performance: An empirical study in germany. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00220-y [GS Search].

Baker, R., Isotani, S., & Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, *19*(02), 03–13. https://doi.org/10.5753/rbie.2011.19.02.03 [GS Search].

Baranyi, M., Nagy, M., & Molontay, R. (2020). Interpretable deep learning for university dropout prediction. *Proceedings of the 21st Annual Conference on Information Technology Education*, 13–19. https://doi.org/10.1145/3368308.3415382 [GS Search].

Barros, T. M., Souza Neto, P. A., Silva, I., & Guedes, L. A. (2019). Predictive Models for Imbalanced Data: A School Dropout Perspective. *Education Sciences*, *9*(4), 275, 1–17. https://doi.org/10.3390/educsci9040275 [GS Search].

Bassetti, E., Conti, A., Panizzi, E., & Tolomei, G. (2022). ISIDE: Proactively Assist University Students at Risk of Dropout. *2022 IEEE International Conference on Big Data (Big Data)*, 1776–1783. https://doi.org/10.1109/BigData55660.2022.10020920 [GS Search].

Beaulac, C., & Rosenthal, J. S. (2019). Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education*, *60*(7), 1048–1064. https://doi.org/10.1007/s11162-019-09546-y [GS Search].

Berka, P., & Marek, L. (2021). Bachelor's degree student dropouts: Who tend to stay and who tend to leave? *Studies in Educational Evaluation*, *70*, 100999. https://doi.org/10.1016/j.stueduc.2021.100999 [GS Search].

Bitencourt, W. A., Silva, D. M., & do Carmo Xavier, G. (2022). Pode a inteligência artificial apoiar ações contra evasão escolar universitária. *Ensaio*, *30*(116), 669–694. https://doi.org/10.1590/S0104-403620220003002854 [GS Search].

Böttcher, A., Thurner, V., Häfner, T., & Hertle, J. (2021). A data science-based approach for identifying counseling needs in first-year students. *2021 IEEE Global Engineering Education Conference (EDUCON)*, 420–429. https://doi.org/10.1109/EDUCON46332.2021.9454042 [GS Search].

Brasil. (1996). *Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas* (tech. rep.). Ministério da Educação, Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras: ANDIFES; ABRUEM; SESu/MEC. Brasília, DF. http://dominiopublico.mec.gov.br/pesquisa/DetalheObraForm.do?select%5C_action=&co%5C_obra=27010 [GS Search].

Chen, Y., Johri, A., & Rangwala, H. (2018). Running out of stem: A comparative study across stem majors of college students at-risk of dropping out early. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 270–279. https://doi.org/10.1145/3170358.3170410 [GS Search].

Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, *96*, 346–353. https://doi.org/10.1016/j.childyouth.2018.11.030 [GS Search].

Colpo, M. P., Primo, T. T., & de Aguiar, M. S. (2021). Predição da evasão estudantil: Uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, 873–884. https://doi.org/10.5753/sbie.2021.218517 [GS Search].

Colpo, M. P., Primo, T. T., Pernas, A. M., & Cechinel, C. (2020). Mineração de dados educacionais na previsão de evasão: Uma RSL sob a perspectiva do congresso brasileiro de informática na educação. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, 1102–1111. https://doi.org/10.5753/cbie.sbie.2020.1102 [GS Search].

Costa, A. G., Mattos, J. C. B., Primo, T. T., Cechinel, C., & Muñoz, R. (2021). Model for prediction of student dropout in a computer science course. *2021 XVI Latin American Conference on Learning Technologies (LACLO)*, 137–143. https://doi.org/10.1109/LACLO54177.2021.00020 [GS Search].

Crespo, C. (2020). Two become one: Improving the targeting of conditional cash transfers with a predictive model of school dropout. *Economia-Journal of the Latin American and Caribbean Economic Association*, *21*(1), 1–45. https://doi.org/10.1353/eco.2020.0011 [GS Search].

da Silva, P. M., Lima, M. N. C. A., Soares, W. L., Silva, I. R. R., de A. Fagundes, R. A., & de Souza, F. F. (2019). Ensemble regression models applied to dropout in higher education. *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, 120–125. https://doi.org/10.1109/BRACIS.2019.00030 [GS Search].

de Assis, B. d. S., Ogasawara, E., Barbastefano, R., & Carvalho, D. (2022). Frequent pattern mining augmented by social network parameters for measuring graduation and dropout time factors: A case study on a production engineering course. *Spcio-Economic Planning Sciences*, *81*. https://doi.org/10.1016/j.seps.2021.101200 [GS Search].

Deho, O. B., Zhan, C., Li, J., Liu, J., Liu, L., & Le, T. D. (2022). How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? *British Journal of Education Technology*, *53*(4), 822–843. https://doi.org/10.1111/bjet.13217 [GS Search].

Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. (2020). Student dropout prediction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12163 LNAI*, 129–140. https://doi.org/10.1007/978-3-030-52237-7_11 [GS Search].

Demeter, E., Dorodchi, M., Al-Hossami, E., Benedict, A., Walker, L. S., & Smail, J. (2022). Predicting first-time-in-college students' degree completion outcomes. *Higher Education*, *84*(3), 589–609. https://doi.org/10.1007/s10734-021-00790-9 [GS Search].

de Oliveira, C. F., Sobral, S. R., Ferreira, M. J., & Moreira, F. (2021). How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review. *Big Data and Cognitive Computing*, *5*(4). https://doi.org/10.3390/bdcc5040064 [GS Search].

Fernández-García, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sánchez-Figueroa, F. (2021). A real-life machine learning experience for predicting university dropout at different stages using academic data. *IEEE Access*, *9*, 133076–133090. https://doi.org/10.1109/ACCESS.2021.3115851 [GS Search].

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). https://doi.org/10.3390/electronics11030457 [GS Search].

Fontana, L., Masci, C., Ieva, F., & Paganoni, A. M. (2021). Performing learning analytics via generalised mixed-effects trees. *Data*, *6*(7). https://doi.org/10.3390/data6070074 [GS Search].

Freitas, F. A. d. S., Vasconcelos, F. F. X., Peixoto, S. A., Hassan, M. M., Dewan, M. A. A., de Albuquerque, V. H. C., & Reboucas Filho, P. P. (2020). IoT System for School Dropout Prediction Using Machine Learning Techniques Based on Socioeconomic Data. *Electronics*, *9*(10), 1613, 1–14. https://doi.org/10.3390/electronics9101613 [GS Search].

Gamao, A., & Gerardo, B. (2019). Prediction-based model for student dropouts using modified mutated firefly algorithm. *International Journal of Advanced Trends in Computer Science and Engineering*, *8*(6), 3461–3469. https://doi.org/10.30534/ijatcse/2019/122862019 [GS Search].

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd). Morgan Kaufmann Publishers. [GS Search].

Hannaford, L., Cheng, X., & Kunes-Connell, M. (2021). Predicting nursing baccalaureate program graduates using machine learning models: A quantitative research study. *Nurse Education Today*, *99*, 104784. https://doi.org/10.1016/j.nedt.2021.104784 [GS Search].

Hoffait, A.-S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*, 1–11. https://doi.org/10.1016/j.dss.2017.05.003 [GS Search].

Hutagaol, N., & Suharjito. (2019). Predictive modelling of student dropout using ensemble classifier method in higher education. *Advances in Science, Technology and Engineering Systems*, *4*(4), 206–211. https://doi.org/10.25046/aj040425 [GS Search].

Iam-On, N., & Boongoen, T. (2017a). Generating descriptive model for student dropout: a review of clustering approach. *Human-Centric Computing and Information Sciences*, *7*, 1, 1–24. https://doi.org/10.1186/s13673-016-0083-0 [GS Search].

Iam-On, N., & Boongoen, T. (2017b). Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics*, *8*(2), 497–510. https://doi.org/10.1007/s13042-015-0341-x [GS Search].

Kang, K., & Wang, S. (2018). Analyze and predict student dropout from online programs. *Proceedings of the 2nd International Conference on Compute and Data Analysis*, 6–12. https://doi.org/10.1145/3193077.3193090 [GS Search].

Karimi-Haghighi, M., Castillo, C., & Hernandez-Leo, D. (2022). A causal inference study on the effects of first year workload on the dropout rate of undergraduates. In M. Rodrigo, N. Matsuda, A. Cristea, & V. Dimitrova (Eds.), *Artificial intelligence in education, pt i* (pp. 15–27, Vol. 13355). https://doi.org/10.1007/978-3-031-11644-5_2 [GS Search].

Kiss, B., Nagy, M., Molontay, R., & Csabay, B. (2019). Predicting dropout using high school and first-semester academic achievement measures. *2019 17th International Conference* *on Emerging eLearning Technologies and Applications (ICETA)*, 383–389. https://doi.org/10.1109/ICETA48886.2019.9040158 [GS Search].

Kitchenham, B. A., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (tech. rep. No. EBSE-2007-01). School of Computer Science and Mathematics, Keele University. Keele, UK. https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf. [GS Search].

Kurniawati, G., & Maulidevi, N. U. (2022). Multivariate sequential modelling for student performance and graduation prediction. *2022 9th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 293–298. https://doi.org/10.1109/ICITACEE55701.2022.9923971 [GS Search].

Kuzilek, J., Zdrahal, Z., & Fuglik, V. (2021). Student success prediction using student exam behaviour. *Future Generation Computer Systems - The International Journal of Escience*, *125*, 661–671. https://doi.org/10.1016/j.future.2021.07.009 [GS Search].

Lee, S., & Chung, J. Y. (2019). The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. *Applied Sciences-Basel*, *9*(15), 3093, 1–14. https://doi.org/10.3390/app9153093 [GS Search].

Lottering, R., Hans, R., & Lall, M. (2020). A model for the identification of students at risk of dropout at a university of technology. *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 1–8. https://doi.org/10.1109/icABCD49160.2020.9183874 [GS Search].

Marques, L. T., Castro, A. F. D., Marques, B. T., Silva, J. C. P., & Queiroz, P. G. G. (2019). Mineração de dados auxiliando na descoberta das causas da evasão escolar: Um mapeamento sistemático da literatura. *Novas Tecnologias na Educação*, *17*(3), 194–203. https://doi.org/10.22456/1679-1916.99470 [GS Search].

Masood, S. W., & Begum, S. A. (2022). Comparison of resampling techniques for imbalanced datasets in student dropout prediction. *2022 IEEE Silchar Subsection Conference (SILCON)*, 1–7. https://doi.org/10.1109/SILCON55242.2022.10028915 [GS Search].

Mduma, N., Kalegele, K., & Machuve, D. (2019a). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal*, *18:14*, 1–10. https://doi.org/10.5334/dsj-2019-014 [GS Search].

Mduma, N., Kalegele, K., & Machuve, D. (2019b). An ensemble predictive model based prototype for student drop-out in secondary schools. *Journal of Information Systems Engineering and Management*, *4*(3). https://doi.org/10.29333/jisem/5893 [GS Search].

Mduma, N., & Machuve, D. (2021). Machine learning model for predicting student dropout: A case of tanzania, kenya and uganda. *2021 IEEE AFRICON*, 1–6. https://doi.org/10.1109/AFRICON51333.2021.9570956 [GS Search].

Nagy, M., & Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, 000389–000394. https://doi.org/https://doi.org/10.1109/INES.2018.8523888 [GS Search].

Naseem, M., Chaudhary, K., & Sharma, B. (2022). Predicting freshmen attrition in computing science using data mining. *Education and Information Technologies*, *27*(7), 9587–9617. https://doi.org/10.1007/s10639-022-11018-3 [GS Search].

Nuanmeesri, S., Poomhiran, L., Chopvitayakun, S., & Kadmateekarun, P. (2022). Improving dropout forecasting during the covid-19 pandemic through feature selection and multilayer perceptron neural network. *International Journal of Information and Education Technology*, *12*(9), 851–857. https://doi.org/10.18178/ijiet.2022.12.9.1693 [GS Search].

Opazo, D., Moreno, S., Alvarez-Miranda, E., & Pereira, J. (2021). Analysis of first-year university student dropout through machine learning models: A comparison between universities. *Mathematics*, *9*(20). https://doi.org/10.3390/math9202599 [GS Search].

Oreshin, S., Filchenkov, A., Petrusha, P., Krasheninnikov, E., Panfilov, A., Glukhov, I., Kaliberda, Y., Masalskiy, D., Serdyukov, A., Kazakovtsev, V., Khlopotov, M., Podo-lenchuk, T., Smetannikov, I., & Kozlova, D. Implementing a machine learning approach to predicting students academic outcomes. In: 2020, 78–83. https://doi.org/10.1145/3437802.3437816 [GS Search].

Orooji, M., & Chen, J. (2019). Predicting louisiana public high school dropout through imbalanced learning techniques. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 456–461. https://doi.org/10.1109/ICMLA.2019.00085 [GS Search].

Ortigosa, A., Carro, R. M., Bravo-Agapito, J., Lizcano, D., Alcolea, J. J., & Blanco, O. (2019). From lab to production lessons learnt and real-life challenges of an early student-dropout prevention system. *IEEE Transactions on Learning Technologies*, *12*(2), 264–277. https://doi.org/https://doi.org/10.1109/TLT.2019.2911608 [GS Search].

Pachas, D. A. G., Garcia-Zanabria, G., Cuadros-Vargas, A. J., Camara-Chavez, G., Poco, J., & Gomez-Nieto, E. (2021). A comparative study of who and when prediction approaches for early identification of university students at dropout risk. *2021 XLVII Latin American Computing Conference (CLEI)*, 1–10. https://doi.org/10.1109/CLEI53233.2021.9640119 [GS Search].

Palacios, C. A., Reyes-Suarez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile. *Entropy*, *23*(4), 485, 1–23. https://doi.org/10.3390/e23040485 [GS Search].

Park, H. S., & Yoo, S. J. (2021). Early dropout prediction in online learning of university using machine learning. *International Journal on Informatics Visualization*, *5*(4), 347–353. https://doi.org/10.30630/JOIV.5.4.732 [GS Search].

Perchinunno, P., Bilancia, M., & Vitale, D. (2021). A statistical analysis of factors affecting higher education dropouts. *Spcial Indicators Research*, *156*(2-3, SI), 341–362. https://doi.org/10.1007/s11205-019-02249-y [GS Search].

Perez, B., Castellanos, C., & Correal, D. (2018). Predicting student drop-out rates using data mining techniques: A case study. In A. Orjuela-Canon, J. Figueroa-Garcia, & J. Arias-Londono (Eds.), *Applications of computational intelligence, ColCACI 2018* (pp. 111–125, Vol. 833). https://doi.org/10.1007/978-3-030-03023-0_10 [GS Search].

Pontili, R., Staduto, J., & Henrique, J. (2018). Abandono e atraso escolar e sua relação com indicadores socioeconômicos: Uma análise para a região sul do brasil. *Gestão & Regionalidade*, *34*(101), 4–22. https://doi.org/10.13037/gr.vol34n101.4173 [GS Search].

Prada, M. Á., Domínguez, M., Vicario, J. L., Alves, P. A. V., Barbu, M., Podpora, M., Spagnolini, U., Pereira, M. J. V., & Vilanova, R. (2020). Educational data mining for tutoring support in higher education: A web-based tool case study in engineering degrees. *IEEE Access*, *8*, 212818–212836. https://doi.org/10.1109/ACCESS.2020.3040858 [GS Search].

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). https://doi.org/10.3390/info13090401 [GS Search].

Queiroga, E. M., Lopes, J. L., Kappel, K., Aguiar, M., Araujo, R. M., Munoz, R., Villarroel, R., & Cechinel, C. (2020). A Learning Analytics Approach to Identify Students at Risk of Dropout: A Case Study with a Technical Distance Education Course. *Applied Sciences-Basel*, *10*(11), 3998. https://doi.org/10.3390/app10113998 [GS Search].

Raschka, S., & Mirjalili, V. (2017). *Python machine learning: Machine learning and deep learning with python, scikit-learn, and tensorflow* (2nd). Packt Publishing. [GS Search].

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), e1355. https://doi.org/10.1002/widm.1355 [GS Search].

Rondado de Sousa, L., Oliveira de Carvalho, V., Penteado, B. E., & Affonso, F. J. A systematic mapping on the use of data mining for the face-to-face school dropout problem. In: In *Proceedings of the 13th international conference on computer supported education - volume 1: Csedu*. INSTICC. SciTePress, 2021, 36–47. ISBN: 978-989-758-502-9. https://doi.org/10.5220/0010476300360047 [GS Search].

Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLOS ONE*, *12*(2), e0171207. https://doi.org/10.1371/journal.pone.0171207 [GS Search].

Santos, G., Belloze, K., Tarrataca, L., Haddad, D., Bordignon, A., & Brandao, D. (2020). Evolvedtree: Analyzing student dropout in universities. *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 173–178. https://doi.org/10.1109/IWSSIP48289.2020.9145203 [GS Search].

Santos, J. R., & Zaboroski, E. (2020). Ensino remoto e pandemia de COVID-19: Desafios e oportunidades de alunos e professores. *Interacções*, *16*(55), 41–57. https://doi.org/10.25755/int.20865 [GS Search].

Segura, M., Mello, J., & Hernandez, A. (2022). Machine learning prediction of university student dropout: Does preference play a key role? *Mathematics*, *10*(18). https://doi.org/10.3390/math10183359 [GS Search].

Shiau, Y. (2020). University dropout prevention through the application of big data. *Proceedings of the 2020 3rd International Conference on Information Management and Management Science*, 1–7. https://doi.org/10.1145/3416028.3416029 [GS Search].

Shilbayeh, S., & Abonamah, A. (2021). Predicting student enrolments and attrition patterns in higher educational institutions using machine learning. *International Arab Journal of Information Technology*, *18*(4), 562–567. https://doi.org/10.34028/18/4/8 [GS Search].

Silva, G. P. d. (2013). Análise de evasão no ensino superior: Uma proposta de diagnóstico de seus determinantes. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, *18*(2), 311–333. https://doi.org/10.1590/S1414-40772013000200005 [GS Search].

Silva Filho, R. L. L., Motejunas, P. R., Hipolito, O., & Lobo, M. B. C. M. (2007). A evasão no ensino superior brasileiro. *Cadernos de Pesquisa*, *37*(132), 641–659. https://doi.org/10.1590/S0100-15742007000300007 [GS Search].

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to predict dropout in university students with machine learning. *2018 IEEE International*

Work Conference on Bioinspired Intelligence (IWOBI), 1–6. https://doi.org/10.1109/IWOBI.2018.8464191 [GS Search].

Sorensen, L. C. (2019). "Big Data" in Educational Administration: An Application for Predicting School Dropout Risk. *Educational Administration Quarterly*, *55*(3), 404–446. https://doi.org/10.1177/0013161X18799439 [GS Search].

Tsai, S.-C., Chen, C.-H., Shiao, Y.-T., Ciou, J.-S., & Wu, T.-N. (2020). Precision education with statistical learning and deep learning: a case study in Taiwan. *International Journal of Educational Technology in Higher Education*, *17*(1), 12. https://doi.org/10.1186/s41239-020-00186-2 [GS Search].

Urbina-Najera, A. B., & Mendez-Ortega, L. A. (2022). Predictive model for taking decision to prevent university dropout. *International Journal of Interactive Multimedia and Artificial Intelligence*, *7*(4), 205–213. https://doi.org/10.9781/ijimai.2022.01.006 [GS Search].

Vasquez Verdugo, J., Gitiaux, X., Ortega, C., & Rangwala, H. (2022). Faired: A systematic fairness analysis approach applied in a higher educational context. *LAK22: 12th International Learning Analytics and Knowledge Conference*, 271–281. https://doi.org/10.1145/3506860.3506902 [GS Search].

Vega, H., Sanez, E., De La Cruz, P., Moquillaza, S., & Pretell, J. (2022). Intelligent system to predict university students dropout. *International journal of online and biomedical engineering*, *18*(7), 27–43. https://doi.org/10.3991/ijoe.v18i07.30195 [GS Search].

Villegas-Ch, W., Palacios-Pacheco, X., & Lujan-Mora, S. (2020). A business intelligence framework for analyzing educational data. *Sustainability*, *12*(14). https://doi.org/10.3390/su12145745 [GS Search].

Viloria, A., Garcia Padilla, J., Vargas-Mercado, C., Hernandez-Palma, H., Orellano Llinas, N., & Arrozola David, M. (2019). Integration of data technology for analyzing university dropout. In E. Shakshuki, A. Yasar, & H. Malik (Eds.), *16th International Conference on Mobile Systems and Pervasive Computing (MOBISPC 2019), 14th International Conference on Future Networks and Communications (FNC-2019), 9TH International Conference on Sustainable Energy Information Technology* (pp. 569–574, Vol. 155). https://doi.org/10.1016/j.procs.2019.08.079 [GS Search].

Xu, Y., & Wilson, K. (2021). Early alert systems during a pandemic: A simulation study on the impact of concept drift. *LAK21: 11th International Learning Analytics and Knowledge Conference*, 504–510. https://doi.org/10.1145/3448139.3448190 [GS Search].

Yang, H., Olson, T. W., & Puder, A. (2021). Analyzing computer science students' performance data to identify impactful curricular changes. *2021 IEEE Frontiers in Education Conference (FIE)*, 1–9. https://doi.org/10.1109/FIE49875.2021.9637474 [GS Search].

Yoo, J. S., Woo, Y.-S., & Park, S. J. (2017). Mining course trajectories of successful and failure students: A case study. *2017 IEEE International Conference on Big Knowledge (ICBK)*, 270–275. https://doi.org/10.1109/ICBK.2017.55 [GS Search].

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes? *Proceedings of the Eighth ACM Conference on Learning @ Scale*, 91–100. https://doi.org/10.1145/3430895.3460139 [GS Search].