Predicting Dropout in Higher Education: a Systematic Review

Jailma Januário da Silva[1] **, Norton Trevisan Roman**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo
Sao Paulo – SP – Brazil

{jailma.januario,norton}@usp.br

Abstract. *In this article, we present a systematic literature review, carried out from February to March 2020, on the application of a machine learning technique to predict student dropout in higher education institutions. Besides describing the protocol followed during our research, which includes the research questions, searched databases and query strings, along with criteria for inclusion and exclusion of articles, we also present our main results, in terms of the attributes used by current research on this theme, along with adopted approaches, specific algorithms, and evalution metrics. The Decision Tree technique is the most used for the construction of models, and accuracy and recall and precision being the most used metric for evaluating models.*

1. Introduction

Student evasion in undergraduate education is a factor that has gained notoriety in recent years, being recurrent in public and private institutions in Brazil. According to [Silveira et al. 2019], academic dropout starts at primary school, moving to secondary school and up to higher education. This feature can, in turn, cause several problems to educational institutions and, consequently, to society itself, given the possible shortage of trained professionals for currently open job positions it may cause [Davok 2016].

Since the loss of a student does not bring any benefit to anyone, it becomes necessary to identify the characteristics presented by groups of students that are at risk of desertion. With this information at hand, institutions can take more assertive actions to prevent these students from evading. Along the years, different computational approaches have been proposed to address this problem, with data analysis techniques being widely used on academic records with some promising results (e.g. [Ahmed and Khan 2019, Silveira et al. 2019, Perez et al. 2018]).

In this article, we present a systematic literature review on this theme, carried out with the objective of highlighting current research related to the use of machine learning and data mining techniques to identify students at risk of dropping out in higher education institutions. Defined as a specific research methodology, developed to gather and evaluate the available evidence related to a topic [Biolchini and Travassos 2005], systematic reviews are useful for the identification, evaluation and interpretation of the available research, specific to some topic, research question, or phenomenon of interest [Kitchenham 2004].

Our goal with this review was then to find out which techniques are used, along with the reasons why they are used, to approach the problem. In doing so, we not only intend to better understand the problem, but also to furnish subsidies for better informed decisions. Considering the particularities of data coming from educational contexts, such

DOI: 10.5753/sbie.2021.217437

as the lack of statistical independence between variables [Baker et al. 2011], for example, some algorithms and data analysis tools may not be appropriate. In this sense, a review that could tell researchers current approaches might be very handy, potentially saving time, effort and financial resources.

Besides, this research also aims at contributing to the Grand Research Challenges in Information Systems in Brazil 2016 – 2026[1], by approaching the Research and Education in Data Science topic within these challenges. It does so by collecting a number of techniques used to address drop-out, presenting them in a single document for the reader. The research was carried out in two scientific bases (Scopus and IEEE explore), resulting in a total of 1,275 articles.

After applying inclusion, exclusion and quality criteria (presented in more detail in Section 2), there still remained 21 studies, which build up the corpus of analysis in this work. As it turned out, features more frequently used in predicting students at risk of dropout are personal/social, academic, demographic and socioeconomic characteristics. Also, popular classification approaches to this end are Decision Trees, Bayes Classification methods, and Logistic Regression.

The rest of this article is organised as follows. Section 2 describes the research protocol followed in the review, which comprises, among other things, searched databases, queries, and criteria for inclusion, exclusion, and quality assessment of fetched publications. Section 3, in turn, presents the results of applying this protocol to the searched databases, while Section 4 discusses our overall findings. Finally, in Section 5 we present our final considerations to this research.

2. Materials and Methods

This research starts with the definition of a Review Protocol, so as to systematise the steps we followed through it. Next, the protocol is followed in the search for articles in the selected databases, observing the inclusion and exclusion criteria defined to this end. Resulting articles are then read and classified according to quality criteria. Finally, the articles with highest scores are selected, and their results are presented. In what follows, each of these steps will be described in more detail.

2.1. Review Protocol

The protocol used in this research was based on the model presented by [Biolchini and Travassos 2005] and [Kitchenham 2004], where it is necessary to establish the research objective, questions to be answered in the retrieved material, inclusion, exclusion and quality criteria, as well as to define the databases to be searched, along with search strings and keywords, previously determined from the results of an exploratory review. In this research, we applied the following protocol:

1. Objective: Identify the different data mining and machine learning techniques used to predict students' dropout risk in higher education institutions.
2. Research questions:
    - What kind of data is used to train and test models?
    - Which approaches or techniques are more frequently applied?

[1] http://www2.sbc.org.br/ce-si/arquivos/GranDSI-BR_Ebook-Final.pdf

- What algorithms are used to implement these techniques?
- What methods or techniques are used to evaluate the adopted approaches?

3. Materials: Articles published in conference proceedings and journals available for download at the searched databases.
4. Searched Databases:
    - SCOPUS: technical and scientific articles from conferences and journals; and
    - IEEE Xplore Digital, Library: articles from conferences and journals, related to computer science, electrical engineering and electronics.
5. Keywords: The keywords used to compose search strings for each database were defined through an initial exploratory search. Selected keywords were: data mining, Educational data mining, predictive models, student dropout, machine learning.
6. Language: English.
7. Search strings: From the set of keywords, the following search strings were built, specifically for each database:
    - *SCOPUS*: (TITLE-ABS-KEY (”Data mining” OR ”educational data mining” OR ”machine learning” OR ”prediction techniques”) AND TITLE-ABS-KEY (”university” OR ”graduate”) AND TITLE-ABS-KEY (”dropout” OR ”evasion”)); and
    - *IEEE Xplore Digital Library*: (((”All Metadata”:”Data mining” OR ”educational data mining” OR ”machine learning” OR ”prediction techniques”) AND ”All Metadata”:”university” OR ”graduate”) AND ”All Metadata”:“dropout” OR “evasion”)))
8. Inclusion criteria:
    - Peer-reviewed articles published since 2015; and
    - Studies describing methods and techniques aimed at predicting student drop-out risk.
9. Exclusion criteria:
    - Studies not related to higher education courses offered in a traditional classroom-based setup;
    - Short conference articles;
    - Repeated studies, in which case only the first source of research will be considered;
    - Studies not related to drop-out prediction in undergraduate courses; and
    - Studies that had not undergone a peer-reviewed process.
10. Quality criteria:
    (a) Does the study present some method, technique or tool to assess its results and approaches?
    (b) Does the study have a well-defined research goal and/or questions based on the related literature?
    (c) Does the study compare its results to those of the related literature?
    (d) Was the study run in a real-life context within the general realm of education management?

2.2. Research Procedure

Using the search strings defined in the Review Protocol, articles were retrieved from each of the selected databases. Each retrieved article had its title and abstract read, and those not containing any of the searched keywords were filtered out. In the sequence, abstracts were read once more, along with each article's introduction section, removing those not meeting at least two of the inclusion criteria, or those contemplating at least one of the exclusion criteria. The introduction section was read so as to confirm that the article described research executed in a higher education institution, as opposed to other educational levels, an information not always available in the documents' abstracts.

The remaining articles were then read in their entirety and their quality was assessed according to the criteria defined in the Protocol. To help in this assessment, articles were assigned scores, reflecting the adopted quality criteria, as shown in Table 1. These scores were then added, so as to assign the article an overall quality value. Next, a threshold score was defined, so as to filter out articles not meeting a minimum quality standard. In this research, the threshold was set to one, ruling articles with scores lower than one out of further consideration.

Table 1. Score associated to each quality criterion

Criterion	Score
a,c,d	0,5
b	1

3. Results

Following the protocol described in Section 2.1, the search was carried out on February 29, 2020 at the IEEEXplore Digital Library, resulting in 1,147 articles, and on March 9, 2020 at SCOPUS, which returned 128 articles. As can be seen in Table 2, SCOPUS returned a total of 128 articles. After filtering out by time period, language and keywords in the article's title and abstract, there still remained 57 articles. Upon their analysis according to the exclusion and inclusion criteria other 47 were removed, there remaining a total of 10 articles, to which the quality criteria scores were applied.

Table 2. Selected articles

Source	Articles	*Selected Articles*
IEEE	1.147	11
SCOPUS	128	10

In the end, a total of 21 articles were selected from both databases. These were then read in their entirety, so as to try to answer the research questions defined in the protocol. Finally, Figure 1 illustrates the distribution of the articles actually considered in our analysis according to their publication year. As can be seen in the figure, even though our search included articles published in 2015, there was no article complying with our selection criteria (inclusion and exclusion) in this year, with 2018 standing out as the year with the highest amount of related articles.

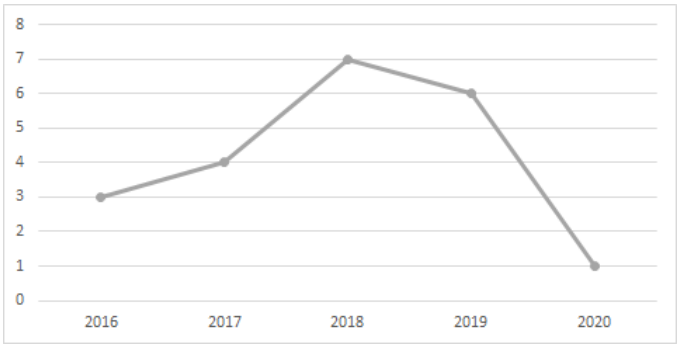

Figure 1. Distribution of articles per publication year.

It is worth recalling, however, that the search was carried out in February and March, 2020, which explains the low amount of articles in this year. In what follows, each of the research questions will be addressed in more detail.

3.1. Question 1: What kinds of data are used to train and test models?

With this question, we sought to determine what types of data and factors are currently being used in computational models to identify students more likely to abandon the course and/or institution, as well as to know the number of attributes that were being analysed to this end. As it turns out, several features can be considered direct contributors to evasion [Davok 2016], whether from the course or the institution, such as:

- Motivational factors: refer to the cause of the student's initial choice for the institution and/or course;
- Psychological factors: refer to issues of autonomy and resilience;
- Institutional factors: refer to university data, such as the number of courses, departments, *campi*, etc.
- Academic features: refer to grades in courses, number of completed courses, amount of failed courses, academic record etc.
- Socioeconomic factors: refer to family income, student assistance, whether the student works or not;
- Demographic factors: refer to the distance from the student's home to the university, if the student came from a different city, state or country;
- Personal/social factors: refer to the student's personal information, such as whether s/he lives alone or shares a home with other students, his/her marital status. etc.
- High-school Type: refers to issues related to the type of secondary school the student attended before entering higher education, along with the method through which s/he entered higher education.

The number of articles, grouped by studied feature, is shown in Figure 2. As it turns out, Personal/Social features (presented in 19 articles – 90,4%), along with Academic features (shown in 16 articles – 76.1%), are the most commonly analysed characteristics in the search for the identification of groups at risk of evasion, followed by Demographic and Socioeconomic features (12 (57,1%) and 9 (42,8%) articles, respectively).

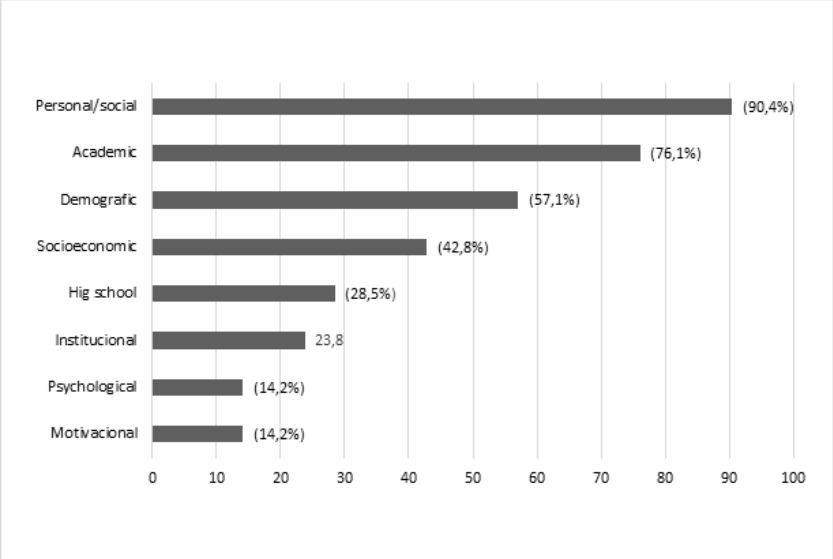

Figure 2. Percentage of articles by analysed feature.

At the other extreme, Psychological and Motivational features are the least analyzed ones, present in only 3 (14,2%) articles. In this analysis, it was also possible to verify that the number of characteristics to be analyzed varied from 11 (*e.g.* [Mayra and Mauricio 2018, Nagy and Molontay 2018, Perez et al. 2018, Iam-On and Boongoen 2017, Hegde and Prageeth 2018]) to 44 (*e.g.* [Timaran Pereira and Caicedo Zambrano 2017, Ahmed and Khan 2019, Srivastava et al. 2019]).

Articles with the lowest amount of features (*e.g.* [Nagy and Molontay 2018, Perez et al. 2018, Iam-On and Boongoen 2017, Hegde and Prageeth 2018]) were those relying on feature selection techniques, such as Principal Component Analysis (PCA), to reduce the dataset's size. The only exception to this rule was that of [Mayra and Mauricio 2018], where researchers had previously defined the number of features to enter their analysis.

In fact, this high number of features considered in each study seems to drive researchers into using different techniques for dimensionality reduction. Nevertheless, when one compares the models built from complete data sets (*i.e.* those to which no dimensionality reduction technique was applied) to those built from reduced data sets, the former perform better than the later.

Regarding the origin of the data sets used in each study, data were usually either retrieved from the educational institution under analysis, as in [Perez et al. 2018], [Timaran Pereira and Caicedo Zambrano 2017],[Solis et al. 2018], [Sivakumar et al. 2016], through questionnaires sent to alumni, as in [Ahmed and Khan 2019],[Hegde 2016].

3.2. Question 2: Which approaches or techniques are more frequently applied?

With this question, we tried to determine which machine learning techniques are more popular amongst researchers, so as to identify good candidates for future research. As it turns out, the majority of the analysed articles implement more than three machine

learning techniques, so as to be able to compare models against each other. Figure 3 shows the fraction of articles implementing each of the identified techniques.

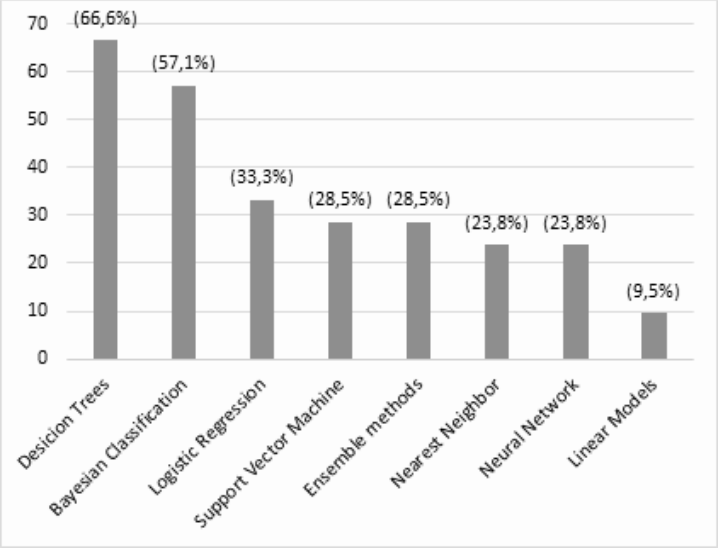

Figure 3. Percentage of articles by identified techniques

As shown in the figure, decision trees stand up as the most popular approach, being present in two thirds of the articles (66.6%, corresponding to 14 articles in total). In the sequence, one finds Bayesian classification methods (more specifically, Naïve Bayes and Bayesian Networks), appearing in 12 articles (57.1% of the total), Logistic Regression (33,3% – 8 articles), SVM and Ensembles (28,5% each – 6 articles) and, finally, Nearest Neighbor and Neural Network models (both Multilayer Perceptron and Deep Neural Networks), with 5 articles each (23.8% of the total) and Linear Models, with 2 articles (9.5%).

3.3. Question 3: What algorithms are used used to implement the techniques?

With this question, we sought to understand which specific algorithms are implemented by the approaches uncovered by Question 2 (Figure 3). Figure 4 illustrates the results. As can be seen in the figure, almost half of the articles implement the Naive Bayes algorithm (47.6% – 10 articles), followed by KNN (K-Nearest Neighbours, with 23.8% – 5 articles) and Random Forests, and the C4.5 algorithm for decision trees (19.4% – 4 articles each).

Bayesian Networks and Generalised Linear Models come next, with 2 articles each (9.5%), with the remaining algorithms being implemented in one (not necessarily the same) article only. It is worth mentioning that 11 articles (52.3%) implementing Decision Trees did not specify the implemented algorithm, as was also the case with 4 articles (19.4%) using Neural Networks. Also, values for Figure 4 do not match those for Figure 3 because some articles implemented more than one algorithm within the same approach.

3.4. Question 4: What methods or techniques are used to evaluate the adopted approaches?

With this question, we tried to understand how the generated models are evaluated, along with what assessment features are taken into account to this end. As shown in Figure 5,

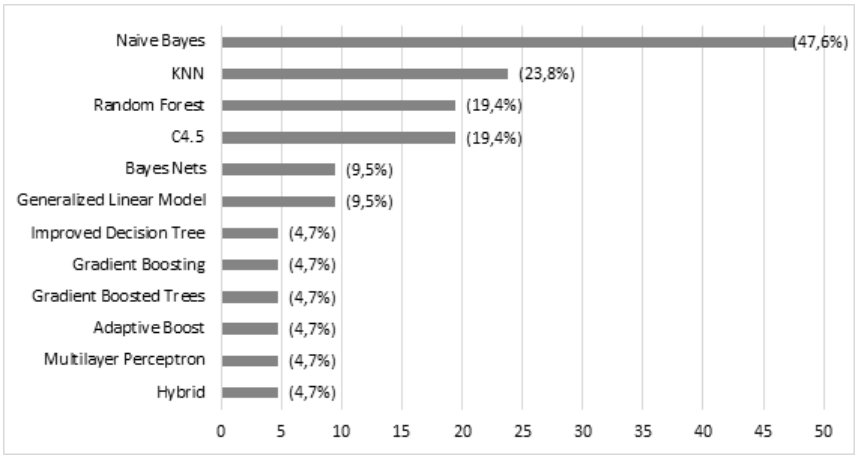

Figure 4. Implemented algorithms

accuracy stands out as the preferred evaluation metric, reported in 42,8% of the articles (9 articles in total). In the sequence one finds recall, with 38% (8 articles), precision, with 33,3% (7 articles), and F1-score, used in 28,5% of the articles (6 articles in total).

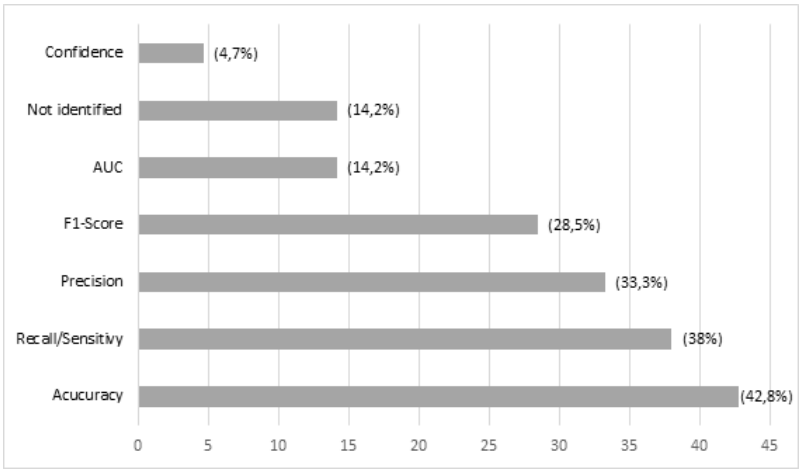

Figure 5. Percentage of studies by evaluation measure

At the other extreme, area under the ROC curve (AUC) was used in 3 articles (14.2%), with a single article (4.7%) presenting confidence intervals for the reported values. Although minority in our research, 14.2% of the articles (3 in total) did not present any evaluation measure for their results. Finally, and as shown in Table 3, it is noticeable that analysed articles majoritarily report more than one evaluation metric. Even though one third of them prefer to stick to a single metric, almost the same amount rely on over three different metrics to evaluate results.

4. Discussion

As it was shown, many different techniques are currently applied to the problem of detecting students at risk of dropout in higher education, ranging from more straightforward ones, such as Naïve Bayes and Logistic Regression, to more elaborate ones, such as Neural Networks and Ensembles methods. As it seems, most of the attention is being paid to

Table 3. Number of different evaluation metrics used.

Articles	*Number of Metrics*
3	0
7	1
4	2
1	3
6	> 3

Decision Trees, perhaps due to their natural intelligibility, even when dealing with a vast amount of data, and for the way they deal with missing values [Gonçalvez 2018]. Along the same lines, methods whose results are harder to interpret, such as Neural Networks, SMVs and Ensembles, come out as less popular choices.

Analysed attributes also seem to be mostly restrained to academic and personal/social features, even though a much wider range of variables is also explored. Concerns about the quality of results also seem to be part of researchers' minds, given the different metrics used to evaluate them, such as accuracy, precision, recall, AUC (Area under the ROC Curve), F1- Score, and even the building of confidence intervals. Nevertheless, some researchers still refrain from carrying out such evaluations, making it harder for the adoption of the proposed solutions.

5. Conclusion

In this article, we described the results of a systematic literature review aimed at identifying current research on the application of machine learning techniques to predict student dropout in higher education institutions. To this end, we presented the protocol we followed when conducting the research, which included the research questions, searched databases, query strings used in each database, along with criteria for inclusion and exclusion of articles.

From the 1,275 articles returned in the database search, the vast majority was discarded for being related to online courses, instead of modalities requiring class attendance. At the end, only 21 articles remained, which constituted our corpus of analysis. Within this set, decision trees are the most frequently applied technique, followed by Bayesian classification and logistic regression. At the end of the scale, we found SVMs, Ensembles, Artificial Neural Networks and Linear Models, along with Nearest Neighbour approaches.

Implementations of these techniques were also found to vary considerably, specially with Ensemble methods, as it would be expected. Also, different evaluations metrics were found to be applied to the implemented techniques, with a predominance of accuracy over others. Even though the vast majority of the articles present their results in terms of such metrics, with most of them presenting more than one metric, we can still find articles where this kind of evaluation was not carried out.

Another interesting result was the focus on persona/social and academic factors to compose the feature vectors used to train models. Despite this preference, other factors are also explored, such as demographic and socioeconomic, amongst others. The number of attributes also varies considerably, from 11 to 44, even though some studies apply

dimension reducing techniques to lower this amount.

Regarding future work directions, it would be interesting to apply the procedure we followed to other databases, so as to compare results in terms of the analysed questions. Such an extensive work might highlight research directions to be explored, and so lead us one step closer to avoid student dropout, something that affects not only public and private educational institutions worldwide, but also society itself.

References

Ahmed, S. A. and Khan, S. I. (2019). A machine learning approach to predict the engineering students at risk of dropout and factors behind: Bangladesh perspective. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6.

Baker, R. S. J. d., Isotani, S., and Carvalho, A. M. J. B. d. (2011). Mineração de dados educacionais: Oportunidadespara o brasil. *Revista Brasileira de Informática na Educação*, pages 1–11.

Biolchini, J., M. P. G. N. A. C. C. and Travassos, G. H. (2005). Systematic review in software engineering. *Systems Engineering and Computer Science Department, UFRJ, Rio de Janeiro*, pages 1–30.

Davok, D. F. Bernard, R. P. (2016). Avaliação dos índices de evasão nos cursos de graduação da universidade do estado de santa catarina – udesc. pages 503–521.

Gonçalvez, T.C. Silva, J. C. C. O. A. (2018). Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do instituto federal do maranhão. *Revista Brasileira de Computação Aplicada*, page 11–20.

Hegde, V. (2016). Dimensionality reduction technique for developing undergraduate student dropout model using principal component analysis through r package. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–6.

Hegde, V. and Prageeth, P. P. (2018). Higher education student dropout prediction and analysis through educational data mining. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pages 694–699.

Iam-On, N. and Boongoen, T. (2017). 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.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Technical Report, Keele University Technical Report TR/SE – 0401, Keele University, Keele, Staffs, UK.*, pages 1–33.

Mayra, A. and Mauricio, D. (2018). Factors to predict dropout at the universities: A case of study in ecuador. In *2018 IEEE Global Engineering Education Conference (EDUCON)*, pages 1238–1242.

Nagy, M. and Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. In *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, pages 389–394.

Perez, B., Castellanos, C., and Correal, D. (2018). Applying data mining techniques to predict student dropout: A case study. *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*, pages 1–6.

Silveira, R. d. F., Victorino, M. d. C., Holanda, M., and Ladeira, M. (2019). Educational data mining: Analysis of drop out of engineering majors at the unb - brazil. *IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1–4.

Sivakumar, S., Venkataraman, S., and Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology*, pages 1–5.

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., and Hernandez, M. (2018). Perspectives to predict dropout in university students with machine learning. *2018 IEEE International Work Conference on Bioinspired Intelligence, IWOBI 2018 - Proceedings*, pages 1–6.

Srivastava, A., Saini, S., and Gupta, D. (2019). Comparison of various machine learning techniques and its uses in different fields. In *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 81–86.

Timaran Pereira, R. and Caicedo Zambrano, J. (2017). Application of decision trees for detection of student dropout profiles. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 528–531.