An Exploratory Analysis on Sociodemographics Features Importance For a Predictive Undergraduate Computing Students Dropout Model

Vitor Gabriel Balsanello[1], **Alinne Corrêa Souza**[1],
Francisco Carlos Monteiro Souza[1], **Thiago Cordeiro Damasceno**[2]

[1]Coordenação de Engenharia de Software
Universidade Tecnológica Federal do Paraná – Dois Vizinhos – PR – Brasil

[2]Instituto de Computação (IC)
Universidade Federal de Alagoas (UFAL) - Maceió – AL – Brazil

vitorbalsanello@alunos.utfpr.edu.br

{alinnesouza, franciscosouza}@utfpr.edu.br, thiago@ic.ufal.br

Abstract. *School dropout is a problem faced by education systems worldwide across various levels of education and institutions. In this context, several strategies are studied and tested to address this problem or at least mitigate it. With the advancement of artificial intelligence (AI), particularly machine learning, a promising opportunity arises to develop robust predictive models capable of accurately identifying complex patterns and anticipating dropout cases. This essay explores the alternatives that some authors have found in using machine learning to prevent school dropout, highlighting and comparing aspects of feature engineering adopted and the most relevant characteristics in the training process. By analyzing case studies and recent research, this essay demonstrates the most important variables and the ones most chosen among researchers to create machine learning models, suggesting which paths are more efficient and faster for new research.*

1. Introduction

School dropout is characterized by the abandonment of a student from an educational institution, leading to absenteeism from classes and adversely affecting their learning process. This issue impacts every year and every stage of Brazilian education, being especially critical in the early years of educational development, as noted by dos Reis (2024). Monitoring and predicting which students are likely to drop out, and to what extent, presents a major challenge. Identifying the primary causes of school dropout and understanding these issues before they manifest provides institutions with valuable tools to address the problem. Various strategies are employed to keep students enrolled and engaged.

As highlighted by Kourkoutas and Hart (2015), in the context of early education, the focus is on providing socio-emotional support to help students cope with issues such as stress, bullying, and anxiety. The goal is to create a positive school environment perceived as safe, welcoming, and supportive, which helps maintain students' engagement

DOI: 10.5753/sbie.2024.242685

and motivation to continue their studies, as emphasized by Freeman et al. (2015). Additionally, educational institutions have adopted the strategy of maintaining regular communication with parents, keeping them informed about students' academic progress and offering resources and support to assist them in supporting their children at home.

When addressing dropout in higher education, the problem becomes more complex, necessitating the development of tailored strategies and mechanisms to mitigate it. Understanding when and how the dropout process occurs, as well as identifying the most significant factors that contribute to it, is crucial for effective intervention. Various social aspects, such as gender and geographical location, are recognized as relevant to this phenomenon. Socioeconomic factors, in particular, can significantly influence a student's likelihood of dropping out of higher education.

As noted by Mduma et al. (2015), the large volume of related data and characteristics suggests that artificial intelligence and machine learning can be valuable tools in mitigating dropout rates. These technologies can offer insights that are not easily identified by traditional methods, playing an important role in addressing school dropout.

This work aims to investigate the significance of gender, race and ethnicity, socioeconomic status, and geographical features in developing machine learning models focused on dropout in computing courses. The primary objective is to evaluate how these characteristics influence dropout training and prediction. The paper is organized as follows: an introduction section covering the problem and key concepts; a related works section reviewing the current state of the art; a methodology section detailing the research methods used; a limitations section outlining the study's constraints; and finally, a conclusion summarizing the study's findings.

2. Background

This section deals with the conceptual aspects and definitions related to the topics covered.

2.1. Higher Education Dropout

School dropout is an extensive and highly complex issue that affects all levels of education systems in various countries. Kehm et al. (2019) defines dropout as a simple concept: The most commonly found description, as cited by the author, suggests that dropout occurs when students leave university before completing their respective courses. Statistically, school dropout can be defined by the contrast between the dropout rate and retention rate or the graduation rate. Although these are the central concepts, school dropout can also be voluntary when the student changes course or university (in which case it is not considered a true dropout).

The primary indicators that suggest a student might drop out or has already dropped out are often linked to their attendance and grades. However, these metrics are more a consequence rather than the root cause of dropout. The main factors are tied to the personal aspects of the student's life, ranging from socio-demographic characteristics to family issues, relationship problems, bullying, and others. These personal issues often precede and significantly contribute to declining academic performance and irregular attendance, which are observable symptoms of a deeper issue.

The main problem resulting from this phenomenon is the hindrance to the student's learning process, as they cannot continue their education. Besides these issues, the economic and social consequences are profound, to the extent that school dropout can be used as an important marker in assessing the quality of life and socio-economic analyses of a country.

2.2. Machine Learning

Awad and Khanna (2015) define machine learning as a branch of artificial intelligence that uses algorithms to uncover patterns in data. Data are raw records or observations, while information is processed data. The field aims to identify relevant patterns in seemingly unrelated datasets using mathematical and statistical methods. Its goal is to generalize training experiences into predictive models based on data behavior.

Currently, the applications of this domain can be seen in various sectors, as reviewed by Angra and Ahuja (2017), from bioinformatics, intrusion detection, information retrieval, game playing, marketing, malware detection, image deconvolution, and more generally, mainly in classification and prediction problems.

2.3. Feature Engineering

Feature engineering is one of the most important parts of the Artificial Intelligence model creation process. According to Zheng and Casari (2018), feature engineering is the activity involved in creating a machine learning model, encompassing the adaptation of raw data in such a way that it becomes useful and usable in the model training process. This is a critical point in the model creation process because selecting and utilizing the most appropriate features can greatly enhance the quality of the final product.

Moreover, it is agreed that a significant portion of the model creation pipeline is dedicated to cleaning and selecting features, both in terms of the time and effort of those involved. An analysis supporting this concept indicates that the correct features can only be chosen in a context that understands both the model and the data, as these may diverge, and it is impossible to generalize the practice of feature selection across different projects.

Currently, feature engineering is a widely studied topic by various professionals, from data scientists to machine learning practitioners, and it plays a crucial role in creating efficient models with relevant accuracy. Therefore, the choice and treatment of features must be done with care and through appropriate techniques, especially in a problem that depends so much on context as the one being addressed.

3. Related Works

This section lists some related works that analyzed and reviewed the strategies, particularly the features adopted by researchers for the school dropout problem, in this case, mainly focusing on the university context. The works related were searched in the Google Scholar Platform. Recent articles, published within the last 5 years and using the same methodology as the current one, such as surveys or essays, addressing the topic were sought.

In Mduma et al. (2019), a series of studies on predicting school dropout were analyzed, listing and summarizing the main strategies and characteristics used by researchers. The conclusions reached by the review show some issues with the datasets used and also a geographic focus, which opens up space for further investigation into the topic.

From another perspective, analyzing online courses this time, but with a similar methodology, Chena et al. (2022), conducted a survey across a wide range of studies. The analyses conducted by the authors indicate that there are interpretability issues in some works, mainly due to the need to understand the unique factors that lead students to drop out of university.

In the year of 2019, Alban and Mauricio (2019), conducted a review of the main data mining techniques used in the problem context. The work revealed that there are a variety of relevant features for student classification, indicating the need for a more specific context to achieve a definitive set of features. The characteristics were grouped into the following categories: personal, academic, economic, social, and institutional factors. The study has showed that, in a general context, a significant proportion of the studies utilized the performance group as the most important characteristic, mainly in relation to college admission test scores.

Oliveira et al. (2021) published a systematic review, in which they evaluated a consistent number of papers on data analysis to identify relevant characteristics in the school dropout process. This study found that the characteristics select by researchers were generally different when considering an overall framework.

The article published by Utomo et al. (2023) used a traditional literature review model, analyzing works from 2018 to 2022. One of the main contributions to the context of the work is that the research revealed the main algorithms used to define the importance of features, such as Correlation Feature Selection, Chi-Square, Information Gain Ratio, and Elastic Net. The article also showed that the correct selection of the feature selection algorithm gradually increases the model's accuracy.

By grouping papers from other authors, the contributions of these articles show that there is a great sparsity of features used in training models for prediction, as well as some gaps in the research, such as the focus on early years and interpretability issues, revealing the complexity of the problem. In addition, the papers generally do not yet include analysis that focuses exclusively on data processing.

4. Investigating Sociodemographic Features for ML Models

This section presents the proposal to investigate and analyze the importance of sociodemographic characteristics such as gender, ethnicity, and location in the development of machine learning models for predicting potential dropouts in computer science courses in higher education. To this end, a methodology was developed that is divided into four steps: (1) definition of a research question; (2) definition of a search strategy; (3) extraction and analysis of the articles; and (4) discussion of the results, as shown in the figure. 1.

Definition of a research question (Step 1): Defining a research question, which will be answered through the search of the analyzed articles. It also helps to ensure that the analysis is focused and targeted and enables the identification of consistent patterns and trends in the analyzed studies. In this case, the following research question is defined:

"RQ_1. Which features have been commonly used for training machine learning models for predicting Higher Education Dropout in Computing Courses?"

Definition of a search strategy (Step 2): Creating a strategy to search articles on

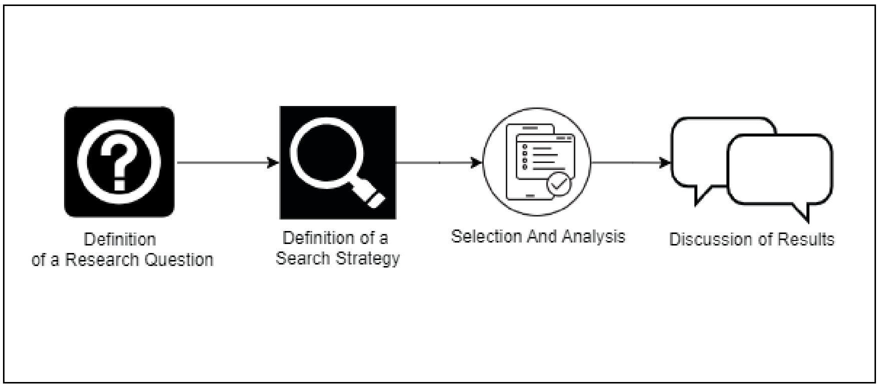

Figure 1. An overview of methodologic process.

the Google Scholar platform related to training machine learning models for predicting Higher Education Dropout in Computing Courses. For this, the following search string is defined:

("machine learning" OR "artificial intelligence") AND ("dropout prediction" OR "student attrition prediction") AND ("Higher Education" OR "university" OR "college")

Selection and Analysis (Step 3): Articles that present the features used and their importance for training the ML model were selected. These features were categorized into three groups: (i) Gender; (ii) Geographic location; and (iii) race/ethnicity.

Only articles that evaluated students in computer science, software engineering, information systems, and related fields were selected, excluding those that evaluated other courses, including these. In this context, the string returned 202 articles, which were then filtered for the last five years, resulting in 161 articles. Among these, 11 articles most related to QP were chosen. Of the 11 articles, two that met the requirements but did not clearly explain or categorize the features were excluded, leaving 9 studies. Four more articles were added by reviewing the references of the previously included eight, totaling 13 articles.

Discussion of Results: The final phase consists of summarizing the results obtained by the researchers and analyzing what these conclusions say about the current state of the problem and how it can be improved.

5. Analysis of Results

This section details the results of the investigation, addressing RQ1. Additionally, a discussion is presented regarding the 13 selected articles and the features defined as gender, geographic location, and race/ethnicity.

5.1. Features used for training machine learning model (RQ1)

As shown in Table 1, the features mentioned in the 13 articles and the accuracy achieved by the related models are summarized.

Shohag and Bakaul (2021) created and compared, including SVM, Decision Tree, Naïve Bayes, KNN, and Random Forest. On average, these algorithms achieved an accuracy of 80%. The features selected for this analysis were more diverse compared to the initial work. The authors considered three aspects: financial information, academic performance, and the students' situation regarding their studies. In this instance, features related to academic skills proved to be particularly significant. The most important features for the model were the Term GPA and the Cumulative GPA (CUM GPA), which are calculated by dividing students' grades by the number of credits for the courses taken. The Term GPA is measured on a quarterly basis, whereas the CUM GPA is measured cumulatively over the entire course.

Table 1. Summary of Features Used for Model Creation

ID	Works Cited	Features	Acurracy
A_1	[Shohag and Bakaul 2021]	Term GPA and Cum GPA	80%
A_2	[Shynarbek et al. 2021]	Grades	96%
A_3	[Ahmed et al. 2021]	Programming Skills and Focus	87%
A_4	[Maksimova et al. 2021]	Grades and Credits	90%
A_5	[Yaacob et al. 2020]	Grades	90%
A_6	[Naseem et al. 2020]	Nationality, age, studying in the home country	80%
A_7	[Dasi and Kanakala 2022]	Access to the university platform, assignments and grades	80%
A_8	[da Silva et al. 2022]	Age and Grades	90%
A_9	[da Cruz et al. 2023]	Class attendance and final grades in subjects	87%
A_{10}	[da Cruz et al. 2024]	Gender and Age	93%
A_{11}	[de O. Santos et al. 2019]	Class Hours	90% and 80%
A_{12}	[Viana et al. 2022]	Disciplines completed within the correct timeframe and average grades	96%
A_{13}	[Dharmawan et al. 2018]	Number of family members, interest in future studies, and relationship with classes	66%

Shynarbek et al. (2021) developed a machine learning model to detect dropouts from a Turkish university, specifically focusing on predicting dropouts from the computer science course. The features used by the authors in this case were the grades in the subjects of the course. The algorithms used for training included Naïve Bayes, Support Vector Machines, Logistic Regression, and Artificial Neural Networks. The analyzed results showed that, in this context, the models could predict the likelihood of students dropping out with significantly high accuracy. However, the size of the dataset may be a relevant issue; the authors analyzed data from 366 students, which may not be sufficient when considering a broader context. In this study, the authors did not include any other features related to the socio-economic context.

Ahmed et al. (2021) conducted a study that resulted in four machine learning models, one based on Decision Trees, SVM, Random Forest, Neural Networks, and Logistic Regression, which achieved accuracies of 81%, 83%, 83%, 83%, and 87%, respectively. In this case, the features used by the researchers were divided into four different groups: Personal and Academic, Analytical Skills, Online and Offline Participation, and Personal Experience and Assessments. In this scenario, the most influential characteristics were, firstly, programming skills, followed by the students' grades in primary and secondary education, and the students' ability to focus, with the latter variable based on a self-assessment conducted by the students themselves. Social aspects were not highly relevant in this case; however, they played a secondary role in model training. Additionally, an important aspect is that the dataset used did not have balanced gender representation, which could be a significant issue, as there were twice as many men as women in the study.

Maksimova et al. (2021) created a machine learning model focused on identifying dropout students in the first year of the computer science course, as the authors identified this period as the most critical. The two best-trained models achieved an accuracy of 90%. In this case, the methodology chosen by the authors involved training models based on different groups of features. The total feature set included data considered by the authors as environmental, such as age, years between school and graduation, and data from the "first-year" group, which included average final grades from the first year, first-year credits, and grades from some subjects. Again, the features that had the most weight in creating the models were the average grade point of the period and the number of credits from the first semester. The other model that achieved a similar score, but with the addition of features indicating whether the student's first year was paid or free, increased the model's accuracy, although this increase was marginal.

Yaacob et al. (2020) conducted a study focused on predicting the likelihood of dropout by considering features related to academic performance data. The features considered were primarily the grades of the courses taken, as well as gender and the weighted average of the final grades based on the credits of the courses. The final model achieved an accuracy of 90% using logistic regression, compared to others trained. In this case, the most important features were grades in Discrete Mathematics, Object-Oriented Programming, Calculus 1, and Data Structures.

Naseem et al. (2020) was one of the few studies to propose incorporating characteristics external to the course. Two machine learning models based on the Random Forest algorithm were created, both achieving an accuracy of 80%. The authors used two evaluation strategies: 5-fold cross-validation and 10-fold cross-validation. Despite the similar accuracy, the second model exhibited higher sensitivity. The authors selected 23 features. Among these, nationality, age, and whether the student is in their home country proved to be quite important, ranking 3rd, 4th, and 5th on the authors' list. Nevertheless, other performance-related characteristics were also significant, with the top three being followed by previous education before college and mathematics grades from the last year of high school.

Dasi and Kanakala (2022) developed a machine learning model by comparing several algorithms: Naive Bayes, Neural Networks, Random Forest, Decision Trees, Logistic Regression, and SVM. In this case, all models exhibited similar performance. The features chosen were filtered from an initial set of 10, based on student behavior. The three features used in the study were the total number of times the student accessed the university's platform, the assignments submitted during the course, and the results of the exams taken.

Da Silva et al. (2022) addressed a university in Portugal and trained a model using a sequence of 23 different features. In this case, the final model achieved an accuracy of 90% using XGBoost, while other models such as CatBoost, ANN, and Random Forest achieved accuracies of 80%. Age was reported as the most important characteristic, with the other 12 features influencing the model being related to student grades in subjects.

da Cruz et al. (2023) developed a scoring system to predict the likelihood of a student dropping out of a computer science course, representing a departure from the approach taken by most studies, which focus on discretely determining whether a student will drop out or not. The features utilized by the researchers encompass various aspects, with significant consideration given to factors such as gender and marital status. In this instance, the features that held the greatest importance for the model were class attendance and final grades in subjects.

In da Cruz et al. (2024) , the authors created a machine learning model by comparing various algorithms, including XGBoost, Decision Tree, and Random Forest. Despite the features not being widely defined, the study managed to identify a clear profile: women between 22 and 28 years old. From this definition, gender and age are considered determining factors in dropout from the course.

O. Santos et al. (2019) demonstrated that one of the important factors for dropping out of the course is related to the semester workload, which can be associated with issues such as work, especially in cases of double workload. The model achieved by the authors attained an accuracy of nearly 90% in IT courses and almost 80% in Computer Science.

In Viana et al. (2022) , a series of classifiers were evaluated to predict undergraduate student dropout, with the analyses yielding results between 85% and 96%. In this study, the authors aimed to generate models for each semester. Several distinct characteristics were used, including grades and failures, while personal and socioeconomic characteristics played a secondary role, appearing in the middle of the ranking of features.

Dharmawan et al. (2018) created a model using non-academic data. The features included information such as gender, distance from home, interest in majors, study motivation, and others. The final model achieved an accuracy of 66% using the random forest algorithm. The most important features, in this case, were, firstly, the number of family members, interest in future studies, and the student's relationship with the class, assessed on three levels: poor, normal, or good. In conclusion, the researchers report that combining these variables with ”internal” variables related to academic performance is necessary to achieve a sufficiently relevant level of accuracy.

5.2. Dataset Analysis

In addition to research focused on the development of machine learning models, there are currently a significant number of publicly available datasets. These datasets are usually published on platforms related to artificial intelligence and data science, such as Kaggle. In general, they cover a wide range of variables, from aspects related to parental education and socioeconomic issues to students' academic performance. This section examines some examples of these datasets.

The criteria for selection included datasets that were returned with the original search string of the studies analyzed, observing the time interval, and differing in their characteristics when compared. In addition, datasets with a large amount of data were preferred.

In Alvarado-Uribe et al. (2019), an extensive dataset of undergraduate students from the University of Monterrey in Mexico was created. The dataset comprises 50 variables and over 140,000 rows of educational data spanning from 2014 to 2020. The information includes variables related to the university, socioeconomic data, and academic information of the students. Regarding socioeconomic variables, these were addressed as follows: the individual's social class was classified into seven levels, gender was analyzed, and ethnicity was not considered by the authors. Additionally, geographic characteristics such as residential neighborhood and whether the student is foreign or not were also considered relevant.

Stein et al. (2022) used a series of characteristics to create a dataset on high school dropout behavior among students in the state of Louisiana in the United States. The dataset contains information from different schools. The variables addressed were created from data from a large number of public schools between the years 2014 and 2019. The dataset includes 86 variables that describe characteristics of school dropout, primarily data on the dropout rate each year, the enrollment rate, and socioeconomic characteristics such as gender and ethnic-racial group.

Valentim Realinho and Martins (2019) created a dataset of students from various universities. The dataset contains information about students from different courses such as agronomy, design, education, nursing, journalism, management, social service, and technologies. Socioeconomic information in this case was represented through characteristics such as marital status, gender, and whether the student has a scholarship, as well as the educational level of the parents. Other characteristics that received considerable attention from the researchers were credits, with the dataset holding information both about the courses passed and the courses taken in the first semester, as well as the courses the student enrolled in.

The dataset constructed by Valdez (2021) combined a series of characteristics, involving academic data. The dataset includes variables that are infrequently included by researchers, such as the number of times the student used the library. In this case, the related socioeconomic variables were the family's social status and gender. Additionally, the dataset shows a series of data related to academic behavior itself, such as absences and grades.

In 2023, the authors Sandra C. Matz et al. (2023) explored a wide range of features in a dataset constructed from information from four universities in the United States, totaling data from 50,095 students. The metrics collected by the authors included both socioeconomic characteristics and data retrieved from institutional applications that reflect student behavior. The authors chose to train models by separating the data by universities. Consequently, features that were most important and appeared in almost all groups were GPA (describing the weighted average of the students' accumulated grades) and ethnic-racial group. These features, although ranked differently, appeared in both the Random Forest and Elastic Net training. Another interesting characteristic is religious identity, which emerged as an important feature for the first university.

5.3. Discussion of Results

Based on the results obtained, one of the first observations is that, although it is well-known that socioeconomic factors directly influence student retention in college, these influences are not always clearly visible in studies. Instead, they often manifest through unexpected variables, such as a decline in students' grades and decreased class attendance. Additionally, it's important to recognize that undergraduate programs can sometimes reflect a social divide from the very start, particularly along gender lines, as men dominate the enrollment in technology and computer-related courses.

Additionally, it is concluded that there is a problem of generalization in the proposed models. Different areas of the dropout problem present unique aspects that require specific variables for each situation. This complexity stems from the unpredictable nature of human behavior, which is influenced by numerous factors, making it difficult to model accurately.

From this perspective, student retention is primarily linked to grades and overall academic performance, both in general terms and when focusing on specific benchmarks. In particular, subjects with a strong mathematical foundation or those fundamental to a student's education, such as data structures, play a significant role in determining performance. Another important factor, mentioned in nearly all studies, is the influence of early education. Approximately 40% of dropouts occur during the first year, and as students progress, models face increasing challenges in predicting the likelihood of dropout.

There are also challenges in modeling and collecting data related to more complex issues. While gender is a relatively straightforward variable, factors such as average family income or mobility difficulties are more complex to capture in datasets and difficult to incorporate into models. These issues are also influenced by other variables that must be considered.

In comparison with similar studies, such as those presented in the related work session and those conducted by Jailma Januário da Silva (2021), Jacobo Roda-Segarra and Mengual-Andrés (2021), and Khalid Oqaidi and Aouhassi (2022), a large part of them focuses on exploring the models themselves rather than the features defined in the datasets, as this study does. Additionally, systematic reviews aim to list the methods and algorithms but do not necessarily discuss the datasets. Table 2 summarizes the results found.

5.4. Analysis of the Defined Features

This section presents an analysis of features such as gender, geographic location, and race/ethnicity based on 13 selected articles. These characteristics are considered key factors in access to and retention in education. Gender affects expectations, opportunities, and learning experiences. Racial issues can systematically reflect inequality and discrimination, and geographic factors reflect the quality and opportunities for access to education as described by Bertocchi and Bozzano (2019), Bonilla-Silva (2021) and Cohen et al. (2021)

Although the studies found did not consider **gender** as a significant issue, there are some aspects that should be taken into account, especially in the Brazilian context.

Table 2. Summary of Results

Aspect	**Description**
Gender Disparity	Male dominance in technology courses reflects a social divide from the start.
Generalization	Models face challenges due to the variability of dropout situations and human behavior.
Academic Performance	Retention is linked to grades and performance in fundamental subjects, such as mathematics and data structures.
Early Dropout	Approximately 40% of dropouts occur in the first year, with increasing difficulty over time.
Data Modeling	Complex data, such as income and mobility, are difficult to model and incorporate.
Comparison with Studies	Studies often focus on models rather than the features of datasets.

The invisible work performed by women, which generally refers to unpaid and often unrecognized tasks carried out at home, in the community, and within the family, can be an important indicator for university dropout. These tasks include various essential activities for household functioning and family well-being, such as childcare, household chores, cooking, shopping, and eldercare, among other responsibilities.

Another aspect that significantly influences student retention in the institution is the issue of **geographic location**. In an initial analysis, students residing in distant locations from the university and needing to commute long distances may face significant costs to continue their studies. Additionally, there is a higher likelihood of dropping out during the course for those residing in very remote areas where commuting is particularly challenging. Other elements to be considered include the availability of public or private transportation, road infrastructure, and student housing options. For students relying on public transportation, the quality and reliability of services can directly affect their ability to attend classes regularly.

Racial/ethnic minorities often face systemic discrimination in education, as seen in Brazil and various studies. This discrimination can range from differential treatment and lack of resources to outright racism and violence, which can demotivate students and increase dropout rates. Additionally, under representation of minorities among teaching staff exacerbates the issue.

It is clear that socioeconomic aspects must be considered in this case. In general, racial and ethnic minorities face socioeconomic disadvantages that can impact their academic performance and their ability to stay in school or university.

6. Limitations of the Study

One of the main limitations of this study is related to the reduced number of datasets specific to the topic, which makes it difficult to explore different analytical approaches.

Additionally, there is inconsistency in the characteristics of each dataset, which can make the comparison between features more complex and affect the generalization of the results obtained.

It is also necessary to highlight the significant contextual differences among the datasets, such as geographic location and social characteristics. Using more similar datasets within a closer context could improve the validity of the results. Another significant challenge is the limited number of studies addressed, reflecting the inherent restrictions imposed by the nature of the topic and the chosen methodology. These factors, combined with issues such as inadequate balancing of characteristics within the datasets and the reduced number of instances in each, can negatively impact the validity and robustness of the conclusions.

However, as a possible future perspective for this work, it is proposed to evolve towards a more comprehensive and complete systematic review, covering the identified gaps and strengthening the existing knowledge base. This approach would not only allow for a deeper and more comparative analysis of the available studies but also the incorporation of new data and methodologies, contributing to a more holistic and detailed understanding of the current topic.

7. Conclusion and Future Works

The problem of university dropout is a significant challenge for modern education, with implications not only for the teaching process but also for social and economic matters. It involves various stakeholders, including parents, teachers, and state institutions. While it may be unavoidable to some extent, dropout can be mitigated. In this context, modern computing strategies can offer new ways to address the issue.

This study analyzes key features for training machine learning models on dropout in computer science courses, highlighting the importance of academic skills and grades. While current research doesn't fully address the specific context, academic performance is identified as the most critical factor. Social and geographical factors influence academic features but do not have a direct impact. Simple machine learning algorithms are sufficient for achieving good classification accuracy.

Most studies on machine learning algorithms in the context of academic dropout highlight grades and attendance as the most important features of the models. However, it is known that students who have dropped out or are at risk of dropping out exhibit these low attributes as a consequence rather than a cause. This study can aid in developing practical strategies to combat dropout, such as creating more accurate predictive models and establishing long-term programs based on longitudinal data analysis. Additionally, it can contribute to enhancing existing models with expanded data. Another valuable perspective would be obtaining feedback directly from students identified as at risk, which could help improve the systems and models.

Future work should focus on developing and validating machine learning algorithms using the sociodemographic characteristics discussed. Given the complex impact of social and geographic variables on student performance and retention, it is crucial to further explore these dimensions. The ultimate goal is to provide educators and policymakers with effective tools to identify at-risk students early and implement targeted interventions.

References

Ahmed, S. A., Billah, M. A., and Khan, S. I. (2021). A machine learning approach to performance and dropout prediction in computer science: Bangladesh perspective. In *16th International Conference on Electronics Computer and Computation*, pages 1–6. IEEE.

Alban, M. and Mauricio, D. (2019). Predicting university dropout through data mining: A systematic literature. In *Indian Journal of Science and Technology*, pages 1–13. Semantic Scholar.

Alvarado-Uribe, J., Mejía-Almada, P., Herrera, A. L. M., Molontay, R., Hilliger, I., Hegde, V., Gallegos, J. E. M., Díaz, R. A. R., and Ceballos, H. G. (2019). Student dataset from tecnologico de monterrey in mexico to predict dropout in higher education. In *Data Mining and Computational Intelligence for E-learning and Education*, pages 1–17. MDPI.

Angra, S. and Ahuja, S. (2017). Machine learning and its applications: A review. In *International Conference on Big Data Analytics and Computational Intelligence*, pages 57–60. IEEE.

Awad, M. and Khanna, R. (2015). *Efficiente Learing Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Apress Open, Índia, first edition.

Bertocchi, G. and Bozzano, M. (2019). Gender gaps in education. *Discussion Paper Series*, pages 1–35.

Bonilla-Silva, E. (2021). *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America*. Rowman and Littlefield, sixth edition.

Chena, J., Fangb, B., Zhangcand, H., and Xue, X. (2022). A systematic review for mooc dropout prediction from theperspective of machine learning. pages 1–14. Taylor e Francis.

Cohen, K., D., and Hill, H. C. (2021). *Learning Policy: When State Education Reform Works*. Rowman and Littlefield, yale university press edition.

da Cruz, R. C., Juliano, R. C., Souza, F. C. M., and Souza, A. C. C. (2023). A score approach to identify the risk of students dropout: an experiment with information systems course. In *Proceedings of the XIX Brazilian Symposium on Information Systems*, pages 1–4. ACM.

da Cruz, R. C., Juliano, R. C., Souza, F. C. M., and Souza, A. C. C. (2024). An exploratory analysis on gender-related dropout students in distance learning higher education using machine learning. In *Proceedings of the XIX Brazilian Symposium on Information Systems*, pages 1–4. ACM.

da Silva, D. E. M., Pires, E. J. S., Reis, A., de Moura Oliveira, P. B., and Barroso, J. (2022). Forecasting students dropout: A utad university study. In *Future Internet*, pages 1–14. MDPI.

Dasi, H. and Kanakala, S. (2022). Student dropout prediction using machine learning techniques. In *International Journal of Intelligent Systems and Applications in Engineering*, pages 1–7. IJISAE.

de O. Santos, K. J., Menezes, A. G., de Carvalho, A. B., and Montesco, C. A. E. (2019). Supervised learning in the context of educational data mining to avoid university students dropout. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, pages 1–2. ACM.

de Oliveira, C. F., Sobral, S. R., Ferreira, M. J., and Moreira, F. (2021). How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review. In *Big Data and Cognitive Computing*, pages 1–33. MDPI.

Dharmawan, T., Ginardi, H., and Munif, A. (2018). Dropout detection using non-academic data. In *4th International Conference on Science and Technology (ICST)*, pages 1–4. MDPI.

dos Reis, R. R. (2024). Políticas públicas de combate a evasão escolar no brasil. Orientador: Marcus Vinícius Costa da Conceição.

Freeman, J., Simonsen, B., McCoach, D. B., Sugai, G., Lombardi, A., and Horner, R. (2015). An analysis of the relationship between implementation of school-wide positive behavior interventions and supports and high school dropout rates. In *The High School Journal*, pages 290–135. The University of North Carolina Press.

Jacobo Roda-Segarra, C. d.-l.-P. and Mengual-Andrés, S. (2024). Effectiveness of artificial intelligence models for predicting school dropout: A meta-analysis. *Multidisciplinary Journal of Educational Research*.

Jailma Januário da Silva, N. T. R. (2021). Predicting dropout in higher education: a systematic review. *X Congresso Brasileiro de Informática na Educação*, pages 1–11.

Kehm, B. M., Larsen, M. R., and Sommersel, H. B. (2019). Student dropout from universities in europe: A review of empirical literature. In *Hungarian Educational Research Journa*, pages 1–18. AK Journals.

Khalid Oqaidi, K. M. and Aouhassi, S. (2022). Towards a students' dropout prediction model in higher education institutions using machine learning algorithms. *nternational Journal of Emerging Technologies in Learning*, pages 1–16.

Kourkoutas, E. and Hart, A. (2015). *Resilience Based Inclusive Models of Students with Social-Emotional and Behavorial Difficulties or Disabilities*. Cambridge Scholars Publishing, Reino Unido, first edition.

Maksimova, N., Dunajeva, O., and Pentel, A. (2021). Predicting first-year computer science students drop-out with machine learning methods: A case study. In *2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering*, pages 1–8. IEEE.

Mduma, N., Kalegele, K., and Machuve, D. (2015). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal*, 18(14):1–10.

Mduma, N., Kalegele, K., and Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. In *Data Science Journal*, pages 1–11. CoCSE.

Naseem, M., Chaudhary, K., Sharma, B., and Lal, A. G. (2020). Using ensemble decision tree model to predict student dropout in computing science. In *Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–8. IEEE.

Sandra C. Matz, Christina S. Bukow, H. P. C. D. and Stachl, A. D. . C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. pages 1–16. Scientific Reports.

Shohag, S. I. and Bakaul, M. (2021). A machine learning approach to detect student dropout at university. In *International Journal of Advanced Trends in Computer Science and Engineering*, pages 1–8. Warse.

Shynarbek, N., Orynbassar, A., Sapazhanov, Y., and Kadyrov, S. (2021). Prediction of student's dropout from a university program. In *16th International Conference on Electronics Computer and Computation*, pages 1–4. IEEE.

Stein, M., Leitner, M., Trepanier, J. C., and Konsoer, K. (2022). A dataset of dropout rates and other school-level variables in louisiana public high schools. page 10. MDPI.

Utomo, A. P., Purwanto, P., and Surarso, B. (2023). Latest algorithms in machine and deep learning methods to predict retention rates and dropout in higher education: A literature review. In *The 8th International Conference on Energy, Environment, Epidemiology and Information System*, pages 1–8. EDP Sciences.

Viana, F. S., Santana, A. M., and de Andrade Lira Rabêlo, R. (2022). Avaliação de classificadores para predição de evasão no ensino superior utilizando janela semestral. In *XI Congresso Brasileiro de Informática na Educação*, pages 908–919. SBC.

Yaacob, W., Sobri, M., Nasir, M., Norshahidi, and Husin, W. (2020). Predicting student drop-out in higher institution using data mining techniques. In *Journal of Physics: Conference Series*, pages 1–14. ICMSCT.

Zheng, A. and Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techinques for Data Scientists*. O'Reilly Media, Estados Unidos da América, first edition.