Predictive Models for University Dropout at UNIRIO

Henrique Soares Rodrigues[1], Laura O. Moraes[1] and Reinaldo Viana Alvares[1]

[1] Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

henrique.rodrigues@edu.unirio.br, laura@uniriotec.br,

reinaldoviana@uniriotec.br

Abstract. *Dropout rates in higher education are a challenge faced by Higher Education Institutions (HEIs), with personal, academic, and social impacts. This research aims to analyze dropout rates in exact sciences courses at UNIRIO using data science and predictive artificial intelligence techniques. Three studies were conducted: a systematic mapping of the literature on AI algorithms applied to dropout rates, a prediction of dropout rates in the Information Systems course with Decision Tree using only academic data, and a broader analysis with Gradient Boosting, including data from the Federal Revenue Service and Annual Report of Social Information (RAIS), to verify the influence of formal employment and entrepreneurship up to the fourth semester. The results indicate that academic performance is the main factor related to dropout rates, while employment and entrepreneurship did not show significant influence.*

1. Introduction

The role of Higher Education Institutions (HEIs) is to provide their graduates with the ability to achieve academic and professional success, contributing to economic and social progress. However, one of the main problems that can occur in HEIs is dropout [Realinho et al. 2022], which is defined as a student permanently abandoning their studies, i.e., excluding temporary enrollment suspensions [Kehm et al. 2019].

According to [Bardagi and Hutz 2005], trying to reduce dropout rates is not only an educational issue but also a political and economic one, as dropout can lead to social and economic losses [Prestes and Fialho 2018, Saccaro et al. 2019].

According to [Von Foerster 2003], First-Order Cybernetics is the study of relationships between entities within a system, such as the educational system, where both the HEI and the student can be classified as entities. An important aspect of such a relationship is the concept of feedback, where the behavior of one entity is influenced by what happens to other entities. Foerster introduces the concept of Second-Order Cybernetics, where each entity is aware of how its own actions influence the system based on feedback of feedback.

In the educational system, the final status of a student (whether they successfully graduated or dropped out) can be seen as feedback to the HEI, which can then give students feedback of that feedback in the form of new educational policies. For HEIs to provide such feedback of feedback in the form of new educational policies and curriculum reforms, it is important for the HEI to be aware of the success or failure data of their students as feedback from the students to the HEIs.

DOI: 10.5753/cbie_estendido.2025.11700

Therefore, this research can be seen as a way to collect feedback from former students to make it possible for the university to give current students feedback of those feedbacks, aiming to minimize dropout rates.

1.1. Problem Statement

According to the Higher Education Census [Brasil 2023], 48% of students who entered federal universities in Brazil in 2015 dropped out. When filtering for exact science courses, such as mathematics, physics, information systems, computer science, among others, this percentage reaches 55%. In the Bachelor's in Information Systems (BSI) program at the Federal University of the State of Rio de Janeiro (UNIRIO), 49.36% of students dropped out between 2000.1 and 2023.1 [Rodrigues et al. 2024a]. With such alarming dropout percentages, it is clear that there is a problem to be addressed in Brazilian higher education.

To help solve this issue, Artificial Intelligence (AI) algorithms have been recognized as potential tools to identify students at risk of dropping out [Tete et al. 2022], enabling HEIs to take actions so that such students remain enrolled and eventually graduate. These actions may include mentoring programs, academic advising, specific pedagogical interventions, workshops to prepare high school students for university life, and even financial support measures when necessary [de la Cruz-Campos et al. 2023].

2. Theoretical Framework

Figure 1 shows a diagram proposed by [Tinto 1975] to represent a student's decision to drop out. The diagram shows that family background, individual characteristics, and elementary and secondary education influence a student's commitment to good academic performance and social interaction with peers. As demonstrated by the literature [Silva and Roman 2021, Tete et al. 2022], poor academic performance is a crucial factor in a student's decision to drop out.

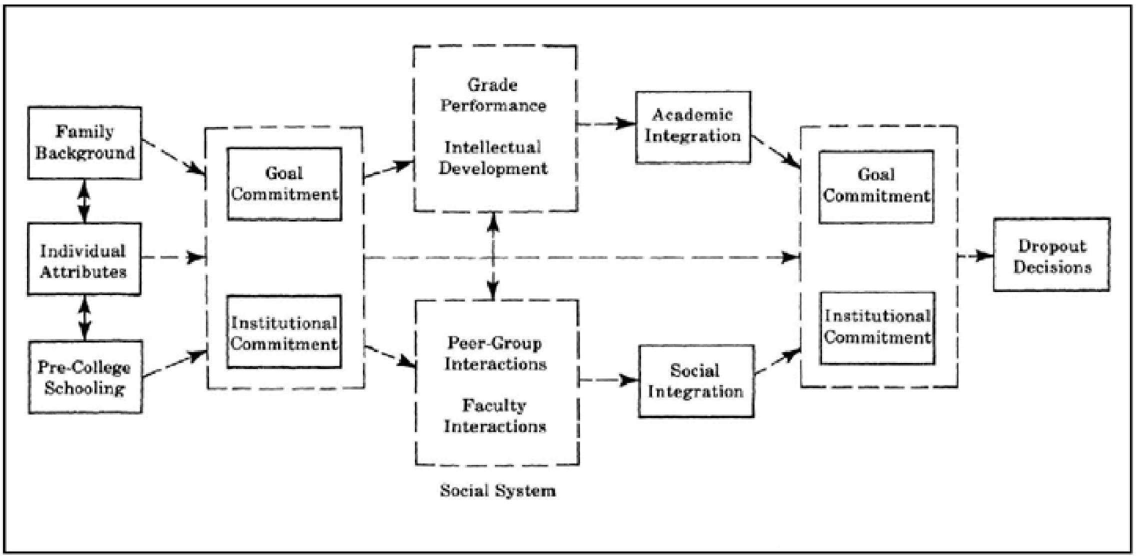

Figure 1. The Decision to Drop Out. Reference: [Tinto 1975]

According to [Samuel 1967], Machine Learning is a subfield of AI that refers to the ability of computers to learn without being explicitly programmed. In this research, predictive models from machine learning, a field of AI, are used to predict student dropout and academic performance. A predictive model is defined by a function:

$$f(X, \beta) = Y$$

where X is the set of predictive features, β represents the unknown factors that may influence the outcome [Baker et al. 2011], and Y is the predicted variable, which in the context of this research can be the final status of a student or their academic performance.

Machine learning models use mathematical functions to predict a target feature, known as y, from other features in the dataset, known as x. Figure 2 represents the machine learning workflow.

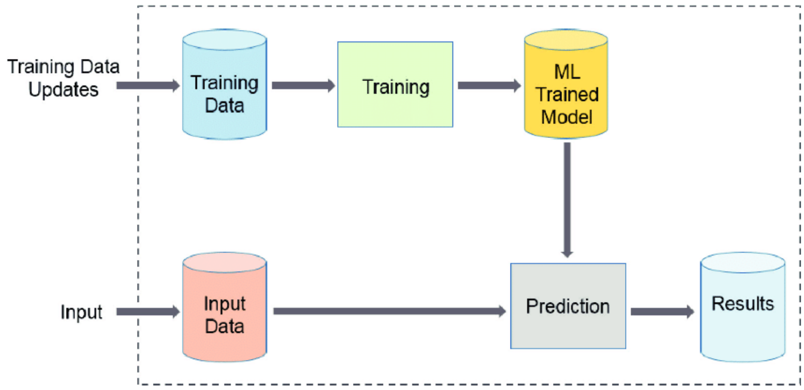

Figure 2. Machine Learning Workflow. Reference: [Osman 2019]

3. Related Work

To obtain the State of the Art, a Systematic Mapping Study (SMS) was conducted following the protocol of [Kitchenham et al. 2023], which is designed for Software Engineering but was adapted for Educational Data Mining (EDM) [Baker et al. 2011]. In summary, our SMS [Rodrigues et al. 2024b] aimed to identify the main algorithms, variables used by researchers worldwide, as well as the difficulties encountered and accuracies achieved by them.

We searched the IEEE Xplore, ACM Digital Library, and Scopus databases using the following search string:

(“higher education” OR “college” OR “graduation” OR “university”) AND (“predict*”) AND (“artificial intelligence” OR “AI” OR “data science” OR “deep learning” OR “machine learning”) AND (“drop off” OR “drop out” OR “dropout”)

We found 223 articles, and after the filtering process, we ended up with 23 articles. As results, we noticed that tree-based algorithms such as Decision Tree, Random Forest, and Gradient Boosting are widely used in the context of dropout prediction. We also saw that the most commonly used variables are related to academic performance and that the main difficulties encountered by researchers are related to data availability, since researchers generally use data from the HEIs to which they are affiliated. The global State of the Art led us to develop a solution prototype presented in section 7.

Also worth mentioning as part of the State of the Art is the dropout prediction feature via logistic regression [Freire et al. 2024] in the Machine Teaching system [Moraes and Pedreira 2020], which uses data from programming exercises in that system to predict which students will not complete basic computing courses. As a result, they found that students who do not practice the exercises are more likely to drop out.

4. Method

4.1. Ethical Issues

This research was approved to use data linkage via the Individual Taxpayer Registry (CPF) by the Ethics Committee under the number CAAE: 78896924.5.0000.5285, complying with the legal conditions established by the General Data Protection Law (LGPD) [Brasil 2018].

4.2. Research Questions

This research aims to answer the following research question (RQ): *"What are the factors and characteristics of academic dropout in UNIRIO's STEM programs?"* To help answer this question, the following subquestions were formulated:

(Sub-RQ1): "What are the most determining variables to predict dropout in UNIRIO's Information Systems program (BSI)?"

(Sub-RQ2): "In which years did the highest dropout rates occur in UNIRIO's BSI program?"

(Sub-RQ3): "Which curricular activities are most decisive for dropout in UNIRIO's BSI program?"

(Sub-RQ4): "Is there an association between full-time employment and dropout rates in CCET?"

(Sub-RQ5): "Is there an association between entrepreneurship and dropout rates in CCET?"

(Sub-RQ6): "Do university scholarships help reduce dropout rates in CCET?"

(Sub-RQ7): "Does a model that includes financial factors along with academic data have more predictive power than a model that considers only academic data?"

5. Dropout Rate Formula

In the dataset, there are several possible values for the student's final status. In this research, students who are currently enrolled, who transferred to another HEI, and those who unfortunately passed away during their studies were excluded.

Students who completed the program successfully and those who dropped out — including abandonment, withdrawal, full cancellation, and dismissal — were considered.

Therefore, considering that Dropped-out Students + Graduates = Total Students, the mathematical formula used to calculate the dropout rate is:

$$\text{Dropout Rate} = \frac{\text{Dropped-out Students}}{\text{Graduates} + \text{Dropped-out Students}}$$

6. Cross-Industry Standard Process for Data Mining (CRISP-DM)

To answer the RQ and Sub-RQs, the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology [Provost and Fawcett 2016], EDM techniques, and the construction of AI models using Gradient Boosting [Friedman 2000] were used.

CRISP-DM was applied in the following phases:

- **Data understanding**: A literature review was conducted to identify which data and models are most commonly used in the topic of dropout prediction with AI. It was observed that financial data are underexplored [Tete et al. 2022, Rodrigues et al. 2024b], so this type of data was chosen to be investigated in this study.
- **Data preparation**: Three data sources were cross-referenced to identify students who became business owners, were employed during their studies, or received scholarships.
- **Modeling**: Exploratory data analysis, chi-square tests, and Gradient Boosting model training were performed.
- **Model evaluation**: The Gradient Boosting models were evaluated.
- **Deployment**: A web system using these models was built.

6.0.1. Chi-Square Test Hypotheses

To determine whether there is a statistically significant relationship between the variables, the chi-square test [Ugoni and Walker 1995] was applied with the following hypotheses:

- H_0 for **sub-RQ4**: there is no statistically significant correlation between dropout and full-time employment.
- H_1 for **sub-RQ4**: there is a statistically significant correlation between dropout and full-time employment.

- H_0 for **sub-RQ5**: there is no statistically significant correlation between dropout and entrepreneurship.
- H_1 for **sub-RQ5**: there is a statistically significant correlation between dropout and entrepreneurship.

- H_0 for **sub-RQ6**: there is no statistically significant correlation between dropout and scholarship receipt.
- H_1 for **sub-RQ6**: there is a statistically significant correlation between dropout and scholarship receipt.

6.1. Model Training Process

Four models were trained for each undergraduate program: Information Systems, Production Engineering, and Mathematics.

All models use financial and sociodemographic data. The first model of each program includes the GPA from the first semester and the grades from first-semester courses. The second model includes the GPA from the first and second semesters and the respective course grades. The third model includes the third-semester GPA, and the fourth model includes all previous data plus the fourth-semester GPA.

General models were also created for the entire CCET. These four models follow a similar structure, but without considering course grades.

To specifically verify sub-RQ7, two models were trained without using financial data to be compared with models that include financial data. This comparison was made using the general CCET models for the first and fourth semesters.

For each of the three programs, four models were created to predict the student's final status. All models use the following variables: admission method, gender, whether the student is an entrepreneur, entrepreneur category, and whether the student is a worker. Each of the four models represents academic progress. Thus, the first model considers the first-semester GPA and grades of the most retained courses in the first semester; the second model includes the GPAs and grades of the two first semesters. These courses were chosen based on the preliminary study results, which showed that poor academic performance impacts dropout. The third and fourth models additionally include the third and fourth semester GPAs, respectively.

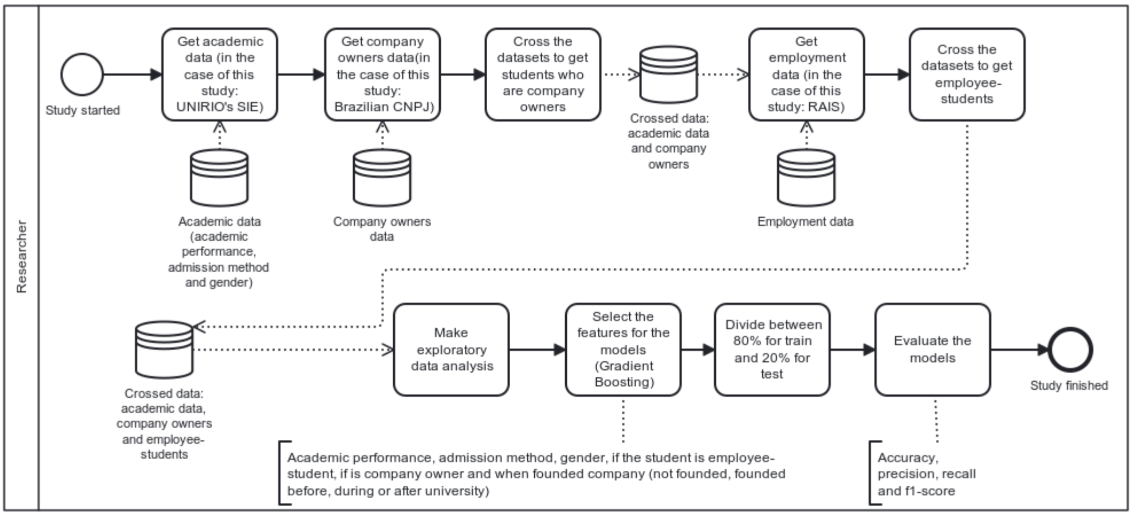

Figure 3. Study process

Figure 3 represents the study process using Business Process Modelling Notation (BPMN) [White 2004]. First, academic data were retrieved from UNIRIO's SIE. Then, this academic data was cross-referenced with CNPJ data to identify which students became business owners. Next, the merged data was again cross-referenced with the RAIS (Annual Social Information Report) data to identify which students worked full-time during their studies. Finally, with the three original datasets merged, data analysis was conducted and the model was trained using the final constructed dataset.

In total, there are 974 unique students, of which 76% are male and 24% female. Students who were still enrolled in the course, who died during the course, or who transferred to other courses or institutions were not included in the dataset. All rows in the dataset were complete, with no missing data. Table 1 presents the dataset after merging.

Table 1. Dataset characteristics

Attribute	Description
Generic student ID	An identifier to link rows to the same student
Program	The student's undergraduate program
Year of enrollment in the curricular activity	The year the student enrolled in the curricular activity
Semester of enrollment in the curricular activity	The semester the student started in the program
Admission method	How the student was admitted (e.g., quotas or open competition)
Final grade in the curricular activity	Final grade the student received in the curricular activity
Gender	Student's gender
GPA in the semester	Student's GPA in the semester
Cumulative GPA	Student's cumulative GPA
Curricular activity name	Name of the curricular activity
Status of the curricular activity	Whether the student passed or failed
Is entrepreneur	Whether the student is a business owner
Entrepreneurship category	Whether the student became an entrepreneur before enrollment, during, or after graduation
Is student-worker	Whether the student had formal employment
Has scholarship	Whether the student received a scholarship
Final status (i.e., whether the student graduated — this is the target variable)	Whether the student completed the program or not

7. Results

In our SMR, we noticed that researchers only published the results of a machine learning experiment using the data they had available, but did not take the next step, which is to build a system that can be used by academic managers to identify students at risk of dropping out. Therefore, we decided to build a system for this purpose using the Gradient Boosting algorithm [Rodrigues et al. 2025]. Figure 4 shows the use cases [Seidl et al. 2015] and Figure 5 shows an example of system usage. The system is available for use by academic managers at http://evasao.uniriotec.br/ and currently contains models for the first four periods of the courses in the Center for Exact Sciences and Technology (CCET), which, in addition to the BSI (Information Systems Bachelor's), also includes Production Engineering and the Mathematics Teaching Degree.

Throughout the development process of this system, we sought to answer research questions about the characteristics of the dropout phenomenon in the BSI and CCET. In the BSI, we found that there is a statistical association between low academic performance and dropout, in agreement with the scientific literature [Silva and Roman 2021, Tete et al. 2022, Rodrigues et al. 2024b], as well as an association with admission

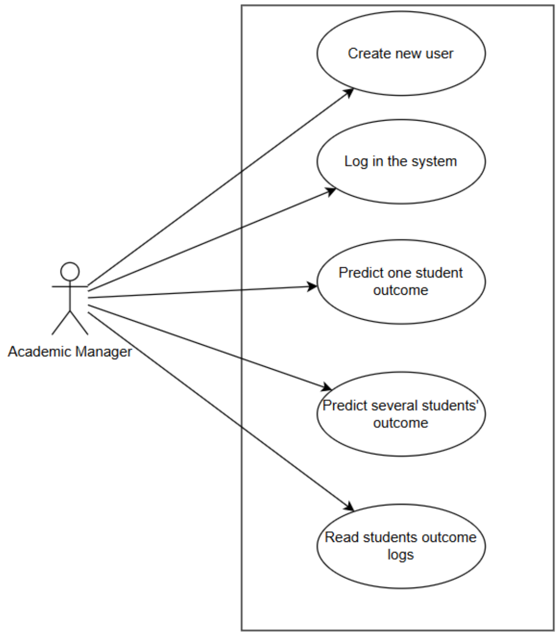

Figure 4. Dropout Prediction System: Use Cases

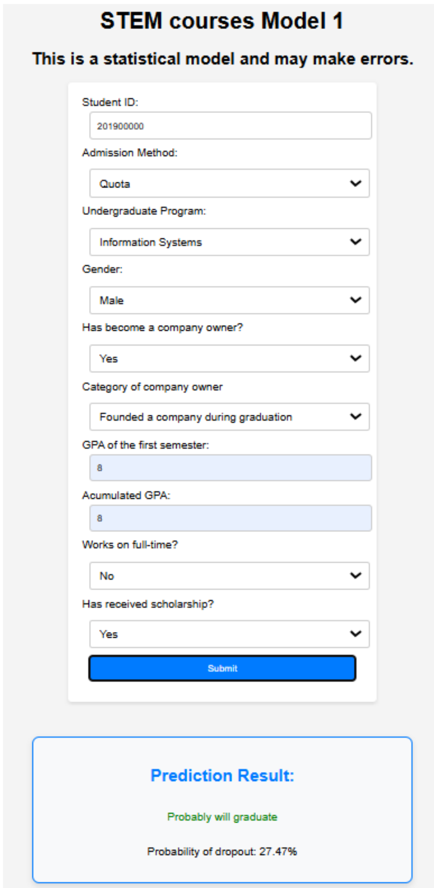

Figure 5. Dropout Prediction System

through the quota system, which was introduced by [Brasil 2012]. We found no statistical association between gender and dropout. Therefore, for **(Sub-RQ1)**, we can answer that the most determinant variables for predicting dropout in the BSI at UNIRIO are those related to academic performance, such as grades and GPA (CR). For **(Sub-RQ2)**, we found that dropouts generally occur in the first part of the BSI course and for **(Sub-RQ3)**, we found that the most decisive curricular activities for dropout in the BSI are those related to basic programming and mathematics.

For CCET, we sought to answer questions involving the student's financial life by cross-referencing UNIRIO data with data from the Ministry of Labor and the Federal Revenue Service, something that is little explored in the literature. This data merging was done through the Individual Taxpayer Registry (CPF), present in all three data sources.

For **(Sub-RQ4)** and **(Sub-RQ5)**, we found no statistical association between dropout and professional activities such as full-time work and entrepreneurship, obtaining p-values of 0.55 and 0.63 respectively in the Chi-Square test. However, since CCET courses are held in the afternoon and evening, such association may exist in morning and full-day courses. For **(Sub-RQ6)**, we found a statistical association with p-value <0.00001 between not receiving scholarships and dropping out, that is, students who receive scholarships graduate in a higher proportion than those who do not.

Figure 6 represents the F1-Score metric for predicting dropouts in the created models. It can be observed that the models become more accurate as the periods progress.

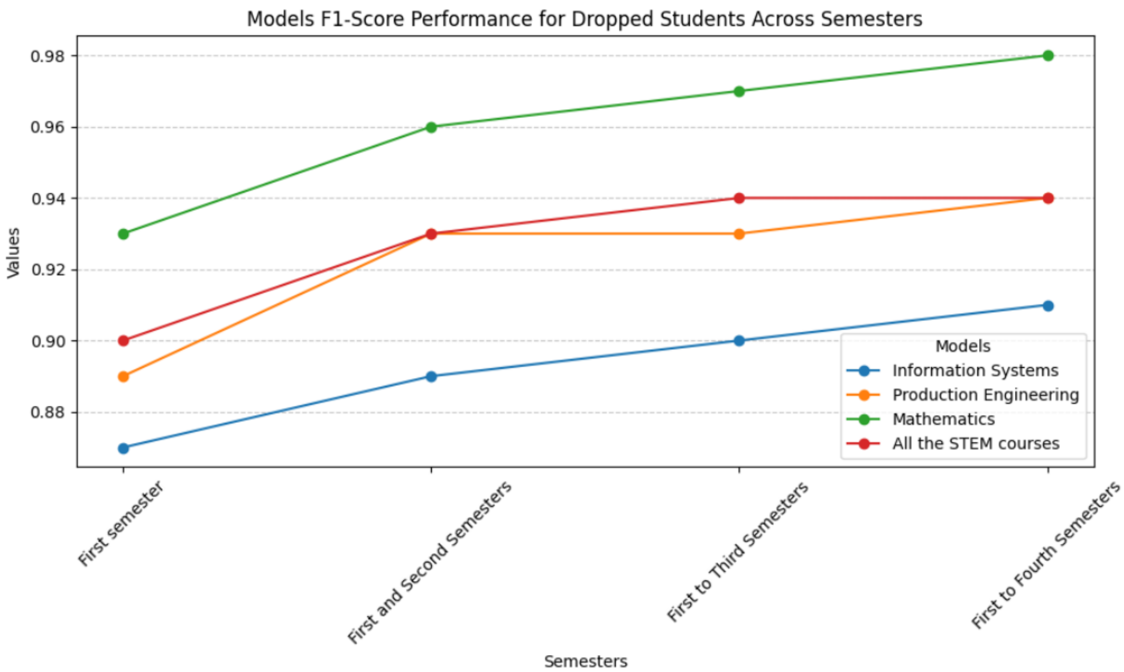

Figure 6. F1 Score for dropped out students in the models

For **(Sub-RQ7)**, we trained a general model for the first period of CCET without financial data to compare with a general model for the first period with financial data. As a result, there was an improvement in the dropout class metrics, especially in recall, by incorporating socioeconomic variables into the models for the first period, while for fourth-period models there was no significant improvement. However, the most predictive socioeconomic variable considered by the model was whether the student receives a scholarship, suggesting that some financial factor may be related to dropout.

Fairness in Machine Learning is defined as the avoidance of bias in a model regarding gender, race, disabilities, and other characteristics [Caton and Haas 2024]. To verify if the models are fair with genders, it was tested the CCET model of the first semester was tested with samples containing only male or female students, exclusively. The model containing only male students had an accuracy of 0.80, and the model containing only female students had an accuracy of 0.82. Although there are more male students in the CCET, this tested model had a higher accuracy in predicting female students' outcomes.

8. Conclusion

The results of this analysis and the models built based on the data from the two main studies demonstrate that students who struggle with curricular activities throughout the course, especially in the first half, are considerably more likely to drop out.

This research found no association between dropout in UNIRIO's STEM courses and financial factors such as professional activities (e.g., full-time work) and owning a business.

It was found that the inclusion of financial data increased the predictive capacity of the model more when combined with academic performance from the first period than with that from the fourth period, indicating that socioeconomic variables are more useful early in the course, before grades from various periods are available.

A system named *Dropout Prediction* was built for UNIRIO's STEM courses, intended to be used by academic managers to identify students at risk of dropping out. It is expected that this system will assist academic managers in reducing dropout rates in UNIRIO's STEM courses. This system was registered in the Brazilian National Institute of Intellectual Property (INPI) in registration number: BR5120250018-4.

This research reinforces the need for higher education institutions like UNIRIO to implement educational policies that help at-risk students remain in their courses and successfully complete their degrees.

8.1. Limitations

This research covered only STEM courses at UNIRIO, so the analysis conclusions may not apply to other contexts such as other UNIRIO courses, other universities, or different geographic locations. Investigating the impact of internships on dropout was considered, but internship data was not available.

The SMS conducted to investigate the use of AI in academic performance prediction showed that previous grades are usually used to predict future ones. This research also had difficulty predicting the GPA (CRA) using factors other than previous academic performance; therefore, no algorithm for predicting academic performance was presented, although that was initially planned.

The *Gradient Boosting* model used in the *Dropout Prediction* system was trained with BSI data from before the curriculum reform implemented in 2023.2. Thus, when used by academic managers from BSI, the system will not reflect the current state of the course. In the future, a new model will be trained according to the updated curricular components of BSI. This issue does not occur in the Production Engineering and Mathematics courses.

Another limitation is that the analysis was restricted to academic and socioeconomic factors. Psychological and health-related factors were not considered.

8.2. Future Work

It is expected that all undergraduate courses at UNIRIO will be included in the *Dropout Prediction* system, as well as the incorporation of factors beyond academic performance and professional life.

Further investigations are needed to understand why the models performed better in predicting dropout among women. An analysis of dropout data based on already completed credits will also be carried out. Courses with a higher weekly workload have more credits.

It is considered future work to use cross-validation in the *Gradient Boosting* models and deploy new models with cross-validation in the *Dropout Prediction* system.

It is also planned to investigate the impact of internships and compare the results of this research with courses that occur only in the morning and afternoon, as there may be dropout differences between students who work and those who do not in such daytime courses.

Future work may also include more studies on the factors that influence academic performance and the construction of an AI model to predict it.

References

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de informática na educação*, 19(02):03.

Bardagi, M. and Hutz, C. S. (2005). Evasão universitária e serviços de apoio ao estudante: uma breve revisão da literatura brasileira. *Psicologia Revista*, 14(2):279–301.

Brasil (2012). Art.14 da portaria mec nº 18/2012. *Diário Oficial da República Federativa do Brasil*.

Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018. *Diário Oficial da República Federativa do Brasil*.

Brasil (2023). Censo da educação superior. https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior/resultados. Acessed in 10/30/2024.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7).

de la Cruz-Campos, J.-C., Victoria-Maldonado, J.-J., Martínez-Domingo, J.-A., and Campos-Soto, M.-N. (2023). Causes of academic dropout in higher education in andalusia and proposals for its prevention at university: A systematic review. In *Frontiers in Education*, volume 8, page 1130952. Frontiers Media SA.

Freire, J., Landim, F., Moraes, L., Delgado, C., and Pedreira, C. (2024). Modelo para previsão precoce de abandono de uma disciplina de introdução à programação. In *Anais do XXXII Workshop sobre Educação em Computação*, pages 635–645, Porto Alegre, RS, Brasil. SBC.

Friedman, J. (2000). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29.

Kehm, B. M., Larsen, M. R., and Sommersel, H. B. (2019). Student dropout from universities in Europe: A review of empirical literature. *Hungarian Educational Research Journal*, 9(2):147 – 164. Place: Budapest, Hungary Publisher: Akadémiai Kiadó.

Kitchenham, B., Madeyski, L., and Budgen, D. (2023). Segress: Software engineering guidelines for reporting secondary studies. *IEEE Transactions on Software Engineering*, 49(3):1273–1298.

Moraes, L. O. and Pedreira, C. E. (2020). Designing an intelligent tutoring system across multiple classes. In *CSEDM@EDM*.

Osman, A. F. (2019). Radiation oncology in the era of big data and machine learning for precision medicine. *Artificial Intelligence-Applications in Medicine and Biology. IntechOpen*, pages 41–70.

Prestes, E. M. d. T. and Fialho, M. G. D. (2018). Evasão na educação superior e gestão institucional: o caso da universidade federal da paraíba. *Ensaio: Avaliação e Políticas Públicas em Educação*, 26:869–889.

Provost, F. and Fawcett, T. (2016). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*.

Realinho, V., Machado, J., Baptista, L., and Martins, M. V. (2022). Predicting student dropout and academic success. *Data*, 7(11).

Rodrigues, H., Moraes, L., and Alvares, R. (2025). Predictive models for university dropout at unirio. Master's thesis.

Rodrigues, H., Moraes, L., Santiago, E., Campos, J., Júnior, E. G., Wanderley, G., Garcia, A., Mello, C., Alvares, R., and Santos, R. (2024a). Predicting student dropout on the information systems undergraduate program of unirio using decision trees. In *Anais do XXXII Workshop sobre Educação em Computação*, pages 588–598, Porto Alegre, RS, Brasil. SBC.

Rodrigues, H., Santiago, E., Wanderley, G., Moraes, L., Eduardo Mello, C., Alvares, R., and Santos, R. (2024b). Artificial intelligence algorithms to predict college students' dropout: A systematic mapping study. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART*, pages 344–351. INSTICC, SciTePress.

Saccaro, A., França, M. T. A., and Jacinto, P. d. A. (2019). Fatores associados à evasão no ensino superior brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de ciência, matemática e computação e de engenharia, produção e construção em instituições públicas e privadas. *Estudos Econômicos (São Paulo)*, 49:337–373.

Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. ii—recent progress. *IBM Journal of research and development*, 11(6):601–617.

Seidl, M., Scholz, M., Huemer, C., and Kappel, G. (2015). *The Use Case Diagram*, pages 23–47. Springer International Publishing, Cham.

Silva, J. and Roman, N. (2021). Predicting dropout in higher education: a systematic review. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 1107–1117, Porto Alegre, RS, Brasil. SBC.

Tete, M. F., Sousa, M. d. M., de Santana, T. S., and Silva, S. F. (2022). Predictive models for higher education dropout: A systematic literature review. *Education Policy Analysis Archives*, 30:(149).

Tinto, V. (1975). Drop-outs from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45:89–125.

Ugoni, A. and Walker, B. F. (1995). The chi square test: an introduction. *COMSIG review*, 4(3):61.

Von Foerster, H. (2003). Ethics and second-order cybernetics. *Understanding understanding: Essays on cybernetics and cognition*, pages 287–304.

White, S. A. (2004). Introduction to bpmn. *Ibm Cooperation*, 2(0):0.