Predicting Student Dropout on the Information Systems Undergraduate Program of UNIRIO Using Decision Tree

Henrique S. Rodrigues[1], Laura O. Moraes[1], Eduardo da Silveira Santiago[1], João Pedro Porto Campos[1], Elmo Sanches Guimarães Júnior[1], Gabriel Monteiro de Castro Xará Wanderley[1], Ana Cristina Bicharra Garcia[1], Carlos Eduardo Ribeiro de Mello[1], Reinaldo Viana Alvares[1], Rodrigo Pereira dos Santos[1]

[1]PPGI - Programa de Pós-Graduação em Informática
UNIRIO - Universidade Federal do Estado do Rio de Janeiro

{henrique.rodrigues, eduardo.santiago, joao.porto,

elmo.junior, gabriel.xara}@edu.unirio.br

{laura, cristina.bicharra, mello, reinaldoviana, rps}@uniriotec.br

Abstract. *This study applied data mining techniques and decision tree algorithm to analyze and predict dropout rates in the Information Systems course at UNIRIO from 2000/1 to 2023/1. Findings show a dropout rate of 49.36%, mostly in the course's first half, with academic performance being a key factor.*

Resumo. *Este estudo aplicou técnicas de mineração de dados e o algoritmo de árvore de decisão para analisar e prever as taxas de evasão no curso de Sistemas de Informação na UNIRIO de 2000/1 a 2023/1. Os resultados mostram uma taxa de evasão de 49,36%, principalmente na primeira metade do curso, sendo o desempenho acadêmico um fator-chave.*

1. Introduction

Higher Education Institutions (HEIs) aim to ensure their students' academic and professional success, contributing to economic growth and social justice. However, student dropout is a significant issue, causing social and economic losses to students, society and HEIs [Prestes and Fialho 2018], also causing the lack of professionals in several areas, compromising an entire necessary ecosystem [Saccaro et al. 2019]. This study defines dropout as students leaving their university studies before completing their degree, not including temporary interruptions, in accordance with Kehm et al. (2019). Bardagi and Hutz (2005) suggested that reducing dropout rates is an educational, economic, and political concern, as it can enhance students' career paths and reduce HEIs' resource wastage. Artificial intelligence algorithms are recognized as valuable tools for addressing student dropouts, as pointed out by literature reviews conducted by Silva and Roman (2021), Tete et al. (2022) and Rodrigues et al. (2024), identifying at-risk students to support their journey to graduation. This study focuses on using artificial intelligence algorithms to predict dropout rates and identify undergraduates before dropping out.

The objective of this study is to identify what corroborates dropout at the Bachelor in Information Systems (BSI) at the Federal University of the State of Rio de Janeiro

(UNIRIO). To do so, data analysis and an artificial intelligence model project were conducted to predict student dropout, thereby helping academic administrators understand this issue from the semester 2000/1 to the semester 2023/1, covering 23 years. This study contributes by providing academic managers and researches with an overview about the dropout issue in a computing undergraduate program that may be have possible similarities in o another undergraduate programs.

The remainder of this paper is structured as follows: Section 2 details the related work; Section 3 presents the planning and conduct of this study; Section 4 explores the results of the data analysis and the decision tree model, as well as a discussion of the findings; Section 5 reports the threats to validity of the study; and Section 6 points out final remarks, limitations, and future work.

2. Related Work

The methodology proposed by Santos et al. (2020) integrates a decision tree with a genetic algorithm and employs cluster-stratified sampling. The findings indicate the need for vigilant monitoring of students with a Grade Point Average (GPA) below 5.79 who have been enrolled for over a year, as they demonstrate a tendency to exceed the stipulated program duration or discontinue their studies. Notably, approximately one-third of the identified dropout cases occurred within the first year of enrollment. The study examines particular attributes, namely ethnicity (race/skin color), participation in social programs, and performance in university entrance exams.

Moseley and Mead (2008) collected two types of data. The first type consisted of time-invariant items, known at the entry point, such as age, gender, educational qualifications, and branch of nursing. The second type comprised time-varying items related to the student's performance, which changed over time. This included semester-wise grades for various modules and attendance records (both gross and net). The data were obtained from 528 students over a five-year entry period, resulting in a total of 3,978 individual records. Using a test set comprising data not seen previously, the system successfully predicted 84% of individuals who subsequently withdrew prematurely. Among those flagged as at risk, 70% indeed withdrew prematurely. When combined, these figures resulted in a high accuracy rate of 94% for the predictions.

Jimenez-Macias et al. (2022) analyzed data from 30,576 students enrolled in a HEI spanning from 2000 to 2020. The findings revealed that variables ralated to GPA, socioeconomic factors, and course pass rates significantly impact the model, irrespective of the semester, faculty, or program. Moreover, a notable disparity in predictive power was observed between Science, Technology, Engineering, and Mathematics (STEM) programs and humanities programs.

3. Research Method

For this study, we used Educational Data Mining (EDM) techiniques [Baker et al. 2011] and data from BSI at UNIRIO, located in Rio de Janeiro, Brazil. This undergraduate course is part of the School of Applied Informatics and enrolls about 30 students per semester. Table 1 summarizes the features collected from the academic system and used in this study. After balancing the dataset based on feature engineering methods, such as removing rows with missing data and irrelevant outliers for the study (i.e., grades above

the allowed average), the database contains 32,382 rows concerning the students' grades in each curricular activity, i.e., the same student appears several times in the database.

At the end, there are 853 distinct students. The dataset consists of students who enrolled from the first semester of 2000 (beginning) to the first semester of 2023. Students who were still enrolled in the course, those who passed away during the course, and those who were transferred to other courses or institutions were not included in the database. All rows in the dataset were filled, there were no problems with missing data. In accordance with the Brazilian General Law on Data Protection (LGPD), the School of Applied Informatics did not disclose the data that compromised anonymity, such as name and identification number [Brasil 2018].

Table 1. Database features

Feature	**Description**
Generic Student ID	An ID to identify rows of the same student
Year of enrollment in curricular activity	The year when the student enrolled in the curricular activity
Semester of enrollment in curricular activity	The first semester when the student enrolled in the course
Admission method to the course	How the student was admitted to the course (i.e., if the student was admitted by quota or broad competition)
Final grade in curricular activity	The final grade the student received in the curricular activity
Gender of student	The student's gender
GPA in the semester	The student's GPA for the semesters
Accumulated GPA of former student	The student's accumulated GPA
Curricular activity name	The name given to each curricular activity of the course.
Status of curricular activity	Whether the student approved or failed the curricular activity
Status of evasion (i.e., course completed or not as this refers to what is predicted) **(target class)**	Whether the student completed the course or not

After performing a brief analysis of the possible artificial intelligence models for this study, we selected Decision Tree, as it was one of the most popular algorithms found in a systematic mapping study conducted in our previous work [Rodrigues et al. 2024]. A decision tree is a tree-structured classification model that can be efficiently induced from the data. Induction of decision trees is one of the oldest and most popular techniques for learning discriminatory models [Fürnkranz 2010]. The model was created using Python on Google Colab.

The goal of this work is to answer the following research question **(RQ)**: *"What contributes to the phenomenon of dropout in BSI at UNIRIO?"* In addition, it seeks to address the following three sub-questions: **(Sub-RQ1)**: "What are the most determining variables to predict dropout in BSI at UNIRIO?"; **(Sub-RQ2)**: "In which years was

there the highest dropout rate in BSI at UNIRIO?"; and **(Sub-RQ3)**: "Which curricular activities are the most decisive for dropout in BSI at UNIRIO?".

To answer these questions, an exploratory data analysis was conducted and a decision tree model focusing on student data was created. In this context, 80% of the data of each individual student were used to train the model while 20% of the data were used to train the model. There were no crossing validation.

4. Results and Discussion

From 853 distinct former students in the database, 432 (50.64%) graduated and 421 (49.36%) dropped out of the course. Of the total, 681 (79.84%) are male and 172 (20.16%) are female. Table 2 shows the specified number and percentage of each gender by who graduated and dropped out. The chi-square statistic was calculated, obtaining a p-value of 0.129. The null hypothesis was that the gender and the student's outcome (graduation or dropout) were independent. Since we did not reject the null hypothesis, we do not have sufficient evidence to state that there is an association between gender and outcome.

Table 2. Graduation or dropout by gender

Graduation or dropout by gender		
Total males	**Graduated males (%)**	**Dropped out males (%)**
681	49.34%	50.66%
Total females	**Graduated females (%)**	**Dropped out females (%)**
172	55.81%	44.19%

The Unified Selection System (SiSU) is a national university entrance exam and was adopted at UNIRIO in 2013 by a Brazilian resolution [Brasil 2012]. Taking into account only the 234 students who entered after the SiSU was implemented, we performed a chi-square test, excluding students who were admitted before the adoption of SiSU, to test the statistical independence between the admission method and dropout, obtaining a p-value of 0.003. The null hypothesis, which was rejected, was that the admission method and student outcome (graduation or dropping out) were independent, suggesting an association between them. Table 3 shows the specified number and percentage of students by admission method of who graduated and dropped out.

Figures 1 to 3 show the accumulated GPA, semester GPA, and curricular activity grade by who graduated and dropped out, respectively. Visual inspection of the box plot allows us to infer statistical significance between the accumulated and semester GPAs of those who graduated and dropped out. However, it cannot be extended to the grades of curricular activities. In this case, dropouts have a larger interquartile range than graduates, comprising the whole grade spectrum. A reason for this behavior could be a difference in the difficulty of curricular activities, which means that some curricular activities may be responsible for "holding back" some students, leading to dropout. We will further investigate this phenomenon in Section 4.3.

Table 3. Graduation or dropout by admission method

Graduation or dropout by gender		
Number of students admitted before SiSU	**Graduated students admitted before SiSU**	**Dropped out students admitted before SiSU**
619	46.05%	53.95%
Number of students admitted by SiSU quotas	**Graduated students admitted by SiSU quotas**	**Dropped out students admitted by SiSU quotas**
95	30.52%	69.48%
Number of students admitted by SiSU (non-quotas)	**Graduated students admitted by SiSU (non-quotas)**	**Dropped out students admitted by SiSU (non-quotas)**
139	49.64%	50.36%

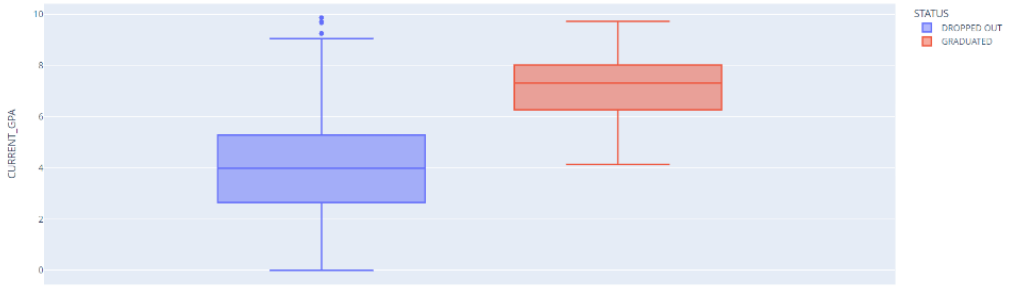

Figure 1. Accumulated GPA by outcome

Figure 4 presents the difference in the average GPA per semester between former students who graduated and those who dropped out. We can notice that after 16 semesters (eight years - the regular course is completed in four years), all dropped out students have already left the university. Those who continue after this period are likely to graduate, taking up to 21 semesters (11 years) to get the degree. In turn, Figure 5 shows the correlation between academic performance features and the outcome status (graduation or dropout). We conducted a Mann-Whitney U Test to verify if graduated and dropped out students perform similar results in CURRENT_GPA, GPA_IN_SEMESTER, and FINAL_GRADE. All testes resulted in a p-value < 0.001, which means that the null hypothesis (two groups of students have the same academic performance) can be rejected.

4.1. Model: Decision Tree

We created a model focused on the student that left the undergraduate program, either successfully or not, using the unique 853 rows and the following features: admission method to the course, accumulated GPA, and gender. These features were used to develop the model because features related to academic performance and socioeconomic conditions were the most common to make this kind of model, according to the literature reviews on this topic [Silva and Roman 2021, Tete et al. 2022, Rodrigues et al. 2024]. Figure 6 presents the result of the decision tree made by the model. The Gini criterion was used and the minimum impurity was defined as 0.005.

As expected from the exploratory data analysis, the model considered the accu-

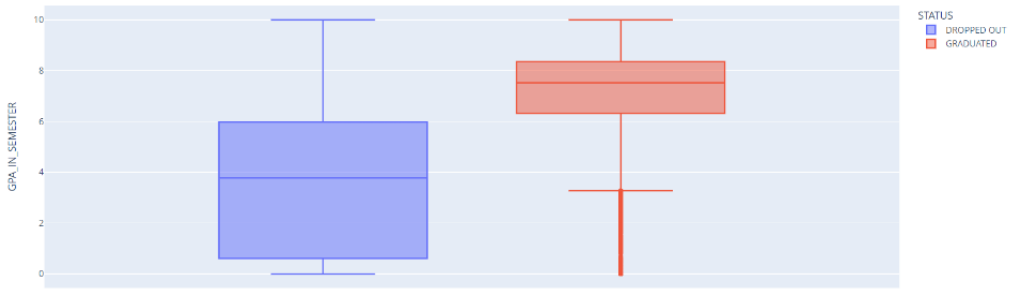

Figure 2. Semester GPA by outcome

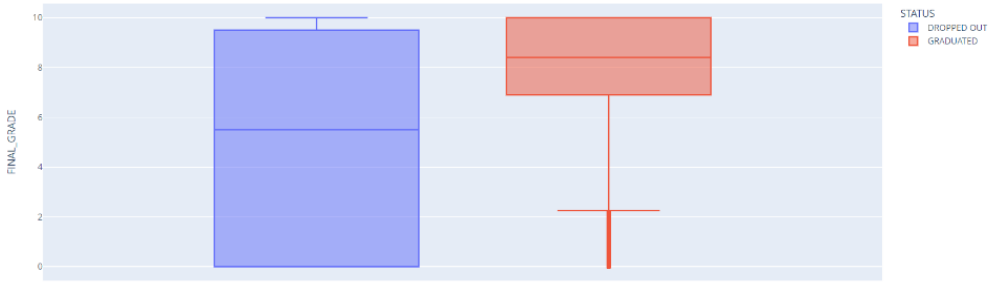

Figure 3. Curricular activities grades by outcome

mulated GPA (in the features named CURRENT_GPA) as the main factor to predict the status of graduation or dropout. The chi-square test suggested an association between the admission method and the predicted outcome. However, this feature and gender were not selected by the model as important features. The model had an accuracy of 83.04%. Table 4 presents the model's results metrics. The model predicted that 83 students graduated, 63 as true graduation and 20 as false graduation. It also predicted that 88 students dropped out, 79 as true dropout and 9 as false dropout.

Table 4. Model accuracy and classification report

Model accuracy and classification report				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.88	0.76	0.81	83
Graduated students	0.80	0.90	0.84	88
accuracy		0.83		

4.2. (Sub-RQ1): What are the most determining variables to predict dropout in BSI at UNIRIO?

After the data analysis and the prediction made by the model, it was demonstrated that the most determining variables to predict university dropout in BSI at UNIRIO is the Accumulated GPA.

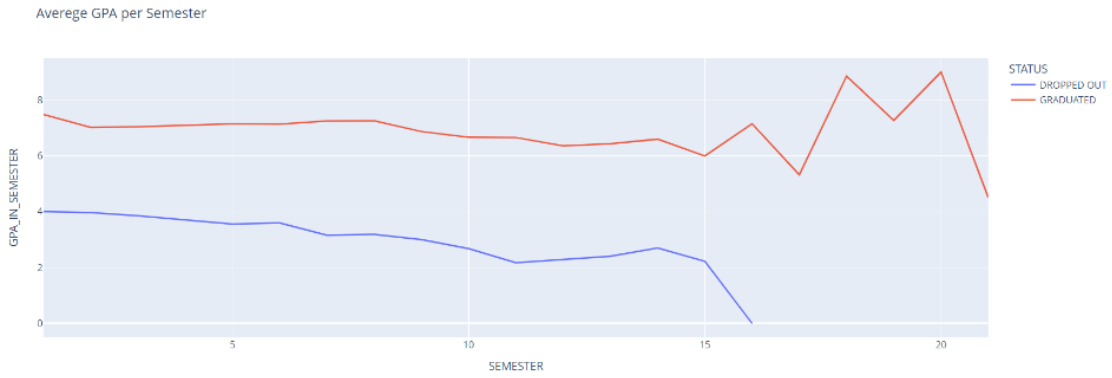

Figure 4. Average semester GPA by outcome

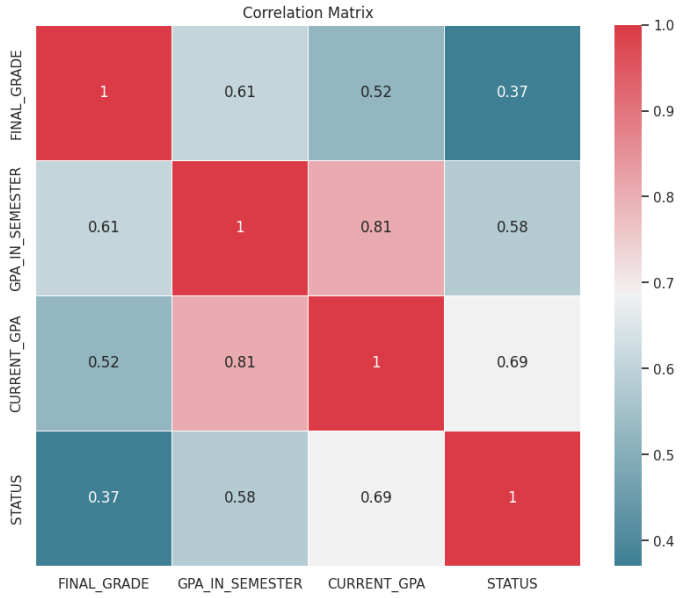

Figure 5. Correlation: academic performance X status (graduation/dropout)

4.3. (Sub-RQ2): In which years was there the highest dropout rate in BSI at UNIRIO?

Figure 7 shows the years until a former students dropped out concretely. Most dropouts occur in the second year of the course, referred as 1 in the graph since it was counted beginning at 0, which comprehends the third and fourth semesters. The fourth year is supposed to be the last year of the course if a student graduates by the established deadline by UNIRIO, but such a stage concentrates the second highest dropout occurrences. The third highest dropouts occur in the first year of the course.

4.4. (Sub-RQ3): Which curricular activities are the most decisive for dropout in BSI at UNIRIO?

Figure 8 shows the top 10 curricular activities that incur the most failures. Academic performance variables are key predictors of university dropout in BSI at UNIRIO. Notably, Accumulated GPA and semester GPA are significant. These GPAs are derived from final grades. Students who fail curricular activities receive grades below 5. Consequently,

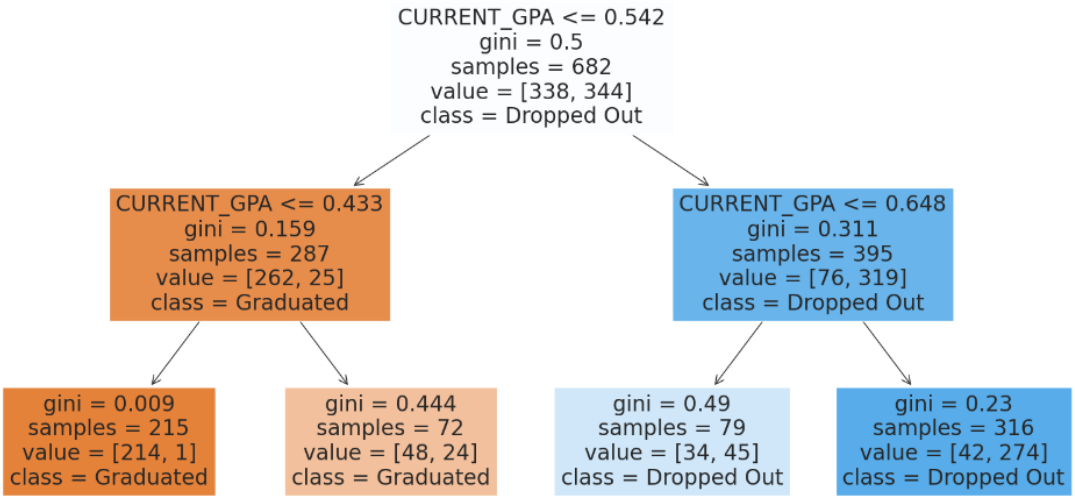

Figure 6. Decision tree of the model focused on the former student

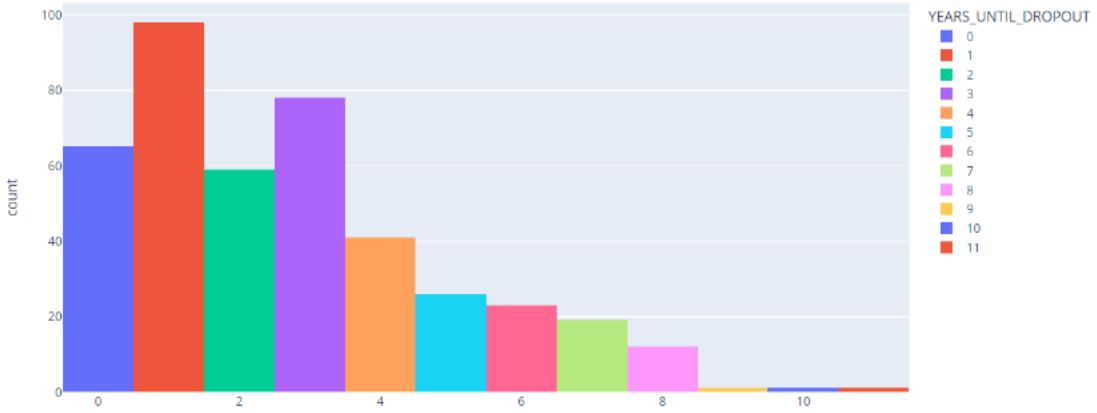

Figure 7. Dropout of students per years since the enrollment

it can be inferred that the curricular activities with the highest failure rates are the most influential in determining dropout in BSI at UNIRIO.

The curricular activity incurring in the most failures is Graduation Project II. This curricular activity is the second step of the writing of the final-year project, which is done at the end of the course. A hypothesis to explain why this curricular activity has many failures is that students can delay the presentation of their final-year project to the following semester, incurring in a failure. A similar hypothesis can be formulated to Extension Curriculum Activities 1 in which students must present university documents of activities they are doing outside, such as internship, courses, or sports. The other eight curricular activities refer to the introduction to programming and mathematics subjects that are concentrated in the first half of the course. Therefore, these curricular activities are the most decisive for university dropout in BSI at UNIRIO.

Most dropouts occurred in the first two years of the course. So, the model can predict early students at risk of dropout by the accumulated GPA of the first two semesters, especially with students who get failures in introduction to programming and mathematics

curricular activities. Schoeffel et al. (2020) found out that students' motivation in introduction to programming can be a indicative of success or dropouts. As such, in a future work, it is possible to verify the students' motivation regarding the BSI's introduction to programming.

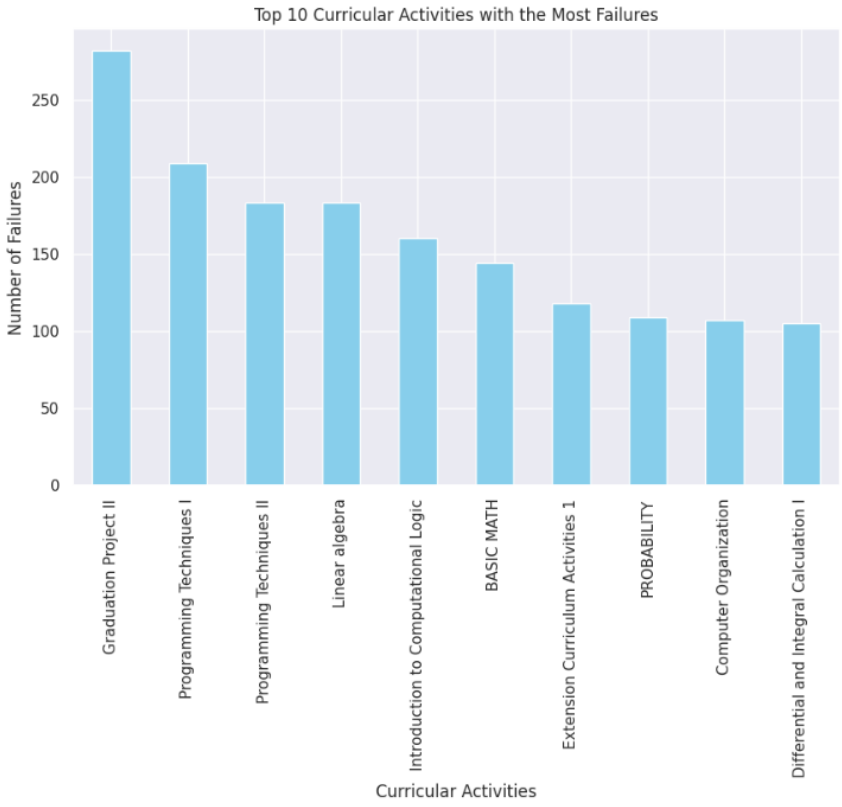

Figure 8. Top 10 curricular activities with the most student's failures

5. Threats to Validity

The main threats to this study are related to data availability. The socioeconomic data available were gender and admission method of former students. Each row of the database is related to a grade obtained by a student on a specific curricular activity, so it was not possible to make a model that agglutinate data from a student, e.g., overall academic performance, such GPA and the final grade on each curricular activities because there would be more instances of GPA unnecessarily, causing a bias on the model. Another threat to validity is that BSI at UNIRIO passed by some curricular reforms throughout the years. As such, the difficulties the students faced have changed over time.

6. Final Remarks

Results from this analysis and the model on data of BSI at UNIRIO shows that students facing difficulties on curricular activities along the course, especially at the first half of the course, have a considerable probability of dropout. This study reinforces the need for HEI, such as UNIRIO, to implement education policies to help students at risk of dropping out to continue their studies and eventually graduate with suscess.

A limitation of this study is that the analysis covers the context of BSI at UNIRIO and the conclusion may not be generalized to other courses at UNIRIO, other universities, and other courses in the field of Computer Science. In future work, one can explore other undergraduate and graduate courses from UNIRIO or other HEIs that agree to make their data available for research. From this study, we hope to help academic administrators with a support for understanding the situation of the dropout problem based on data mining and analysis.

Acknowledgments

This research was supported by CAPES (Financial Code 001), CNPq (Proc. 316510/2023-8), FAPERJ (Procs. 210.688/2019 and 211.583/2019), and UNIRIO (PPQ 2023 and PPInst 2023).

References

[Baker et al. 2011] 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 and Hutz 2005] 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] Brasil (2012). Art.14 da portaria mec nº 18/2012. *Diário Oficial da República Federativa do Brasil*.

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

[Fürnkranz 2010] Fürnkranz, J. (2010). *Decision Tree*, pages 263–267. Springer US, Boston, MA.

[Jiménez-Macias et al. 2023] Jiménez-Macias, A., Moreno-Marcos, P. M., Muñoz-Merino, P. J., Ortiz-Rojas, M., and Kloos, C. D. (2023). Analyzing feature importance for a predictive undergraduate student dropout model. *Computer Science and Information Systems*, 20(1):175–194.

[Kehm et al. 2019] 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.

[Moseley and Mead 2008] Moseley, L. G. and Mead, D. M. (2008). Predicting who will drop out of nursing courses: A machine learning exercise. *Nurse Education Today*, 28(4):469–475.

[Prestes and Fialho 2018] 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.

[Rodrigues. et al. 2024] Rodrigues., H., Santiago., E., Wanderley., G., Moraes., L., Eduardo Mello., C., Alvares., R., and Santos., R. (2024). 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 et al. 2019] 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.

[Santos et al. 2020] Santos, G., Belloze, K., Tarrataca, L., Haddad, D., Bordignon, A., and Brandao, D. (2020). Evolvedtree: Analyzing student dropout in universities. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 173–178.

[Schoeffel et al. 2020] Schoeffel, P., Ramos, V. F. C., and Wazlawick, R. S. (2020). A method to predict at-risk students in introductory computing courses based on motivation. In *Anais do 9º Concurso Alexandre Direne de Teses de Doutorado - Congresso Brasileiro de Informática na Educação (CBIE)*, pages 41–41, Porto Alegre. Sociedade Brasileira de Computação.

[Silva and Roman 2021] 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 et al. 2022] 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).