A Predictive Model for Dropout Risk in a Computer Science Education Program

Marcos V. O. Assis[1], **Anderson S. Marcolino**[1]

[1]Universidade Federal do Paraná – Departamento de Engenharias e Exatas – Palotina – PR – Brasil.

Abstract. *The ever-growing demand for computing professionals requires the effective management of educational resources. With the increasing importance of computer science education programs in Brazil, identifying potential dropout students has become crucial for educational institutions. However, predicting which students are likely to drop out poses a significant challenge, especially in non-metropolitan areas. To address this issue in the Computer Science Education program of the Federal University of Paraná (Brazil), we propose an approach that leverages machine learning to analyze different features associated with the student's academic performance and detect possible dropouts. We compare the performance of 15 machine learning algorithms in predicting student dropouts, additionally identifying the most influential variables contributing to this situation. To evaluate the effectiveness of our approach, we conduct experiments using real data collected from the computer science education program. The results demonstrate the efficacy of our approach in identifying students at risk of dropping out.*

1. Introduction

Nowadays, computing education has become essential, prompting educational institutions to integrate computer and technology disciplines into curricula. In 2019, the SBC (Brazilian Computer Society) proposed guidelines outlining the competencies and skills related to Computer Science. In 2022, Computer Science was integrated into the national standards (BNCC), mandating the inclusion of computing in schools throughout the entire country [BRASIL 2022].

Introducing computer education in basic education is expected to increase demand for Bachelor in Computer Science Education programs [de Almeida and Mateus 2015, Linhares and Santos 2021]. However, no legally established positions exist for such graduates in Basic Education. Currently, teachers without computing backgrounds occupy these roles. This situation leads to low demand for these programs, with only 88 active ones in Brazil (e-MEC)[1]. As a consequence, summed up with the end of the COVID-19 pandemic period, the dropout rate is on the rise [Moscoviz et al. 2022, Colpo et al. 2021].

The historical series from 2010 to 2020, encompassing higher education indicators in Brazil, indicates positive results until 2019. When comparing 2019 with 2020, there was a 9.41% reduction in enrollments across the entire system and a 6.2% reduction in face-to-face courses at public institutions. Regarding graduates of a higher education program, there was a reduction of 22.1% in face-to-face courses at public institutions and 6% across the entire system [Wegner 2022].

[1]`https://emec.mec.gov.br/`

DOI: 10.5753/sbie.2024.242106

According to [Olmedo-Cifuentes and Martínez-León 2022] and [Santos et al. 2021], dropout intentions can be a warning signal. Among those signs are noticed absenteeism, submission of work late or not taking valuation tests (performance), negative class experiences, and a lack of motivation, satisfaction, or commitment. According to an OECD (Organisation for Economic Cooperation and Development) [Nascimento and Verhine 2017], the average annual cost per undergraduate enrolled in the Brazilian higher education system was US$ 13,539.90. The 2017 census on higher education [Fluminense 2015] also showed the problem, with the number of dropped students reaching 1,818,838 [Santos et al. 2020]. Therefore, we can assume that the cost would be around US$ 24 billion for students who did not complete their studies.

The dropout issue is particularly pronounced in campuses located in small cities [Barbosa-Camargo et al. 2021, Lewine et al. 2021]. Many students in these programs come from smaller neighboring towns, facing challenges such as long commuting distances and difficulty in balancing daily on-site activities with their local commitments. This reality, summed with socioeconomic and academic factors, may contribute to the higher dropout rates in these programs. In this scenario, it becomes crucial to understand the elements that influence student attrition to implement effective preventive strategies and actions.

In this study, we introduce an approach that employs machine learning to analyze various factors linked to student's academic performance and identify potential dropouts, taking into account the unique characteristics of the Bachelor's Degree program in Computer Science Education located in the city of Palotina, a small town of 30,000 inhabitants in the countryside of Paraná, in a satellite campus of the Federal University of Paraná (UFPR). Our methodology encompasses the gathering and preprocessing an extensive real dataset of the program, which incorporates student performance in specific courses. Subsequently, we undertake a comparative analysis of the predictive capacities of 15 machine learning algorithms for student abandonment rates while simultaneously identifying the crucial variables that significantly influence this situation.

The main contributions of this manuscript are (i) the contextualization of the Dropout problem of the Bachelor of Computer Science Education program from a satellite campus located in a non-metropolitan area of a Brazilian federal university, providing a data analysis that enables a deeper understanding of the problem; (ii) collected and preprocessed a comprehensive real-world dataset, encompassing student performance features; (iii) compared 15 machine learning algorithms for classifying students as potential dropouts versus non-dropouts; (iv) XGBoost classifier achieved the best performance in identifying dropout students, with 100% recall; (v) the Feature selection process identified the most predictive attributes regarding students' background and academic performance; and the (vi) validation on known potential dropout cases demonstrated the efficacy of the tuned XGBoost model in recognizing high-risk students.

The remainder of the paper is organized as follows. Section 2 delves into related works that address dropout and its prediction or analysis using machine learning techniques. Section 3 introduces the context of the evaluated dropout scenario, detailing the dataset preparation and implementation of machine learning models. Section 4 outlines the results, and Section 5 provides the conclusion.

2. Related Works

As this research focuses on investigating dropout rates in the Bachelor of Computer Science Education program, relevant related works were initially reviewed to align with this context. However, no studies specifically addressing this target undergraduate program were found in repositories such as Springer, IEEE, ACM, and Google Scholar. Consequently, primary studies that examine dropout rates in broader Computer Science undergraduate programs were considered, along with those that employ machine learning techniques to develop dropout analysis and prediction models in the context of on-campus undergraduate Computer Science programs.

[Schefer-Wenzl et al. 2024] conducted a literature survey and qualitative interviews to identify the underlying of dropout in a Computer Science Bachelor´s degree program at the University of Italy. Authors reveal a range of reasons, with time constraints and misaligned expectations of the degree program emerging as the most frequently mentioned factors in their interviews.

In [Bravo et al. 2023], authors aimed to predict student dropout rates in a Bachelor's Degree program in Computer Science at a Brazilian public university, addressing the significant dropout rate exceeding 50% after the fifth year. To achieve this, the researchers employed machine learning techniques, specifically using seven classifiers, including Gradient Boosting (GB), to analyze four distinct datasets representing different academic stages. The results indicated that the most influential features for predicting dropout were the final grades in key disciplines, with the GB algorithm achieving the highest performance, which identified 91% of students likely to drop out.

[Mathews de et al. 2023] aimed to investigate dropout rates among Computer Science undergraduate students at the University of Brasilia (UnB). Utilizing a dataset of 879 observations, the researchers applied survival analysis, specifically fitting a Log-Normal regression model. The results indicated that students with higher grade point averages and those who took summer courses were less likely to drop out. Additionally, students admitted through the National High School Exam (ENEM) and other UnB-specific admission modes demonstrated lower dropout rates than alternative admission pathways.

[Varga and Ádám Sátán 2021] aimed to identify pre-enrollment attributes of first-year Computer Science students at the University of Hungary that contribute to failure in the first-semester Programming Basics course and ultimately to dropout, intending to detect at-risk students early and offer appropriate mentorship. The analysis is based on secondary school performance data, including school rank, admission scores, foreign language knowledge, and first-semester Programming Basics course. The study found that admission scores and school rank significantly impact first-semester Programming Basics results. Additionally, students with weaknesses in all examined pre-enrollment attributes are more likely to drop out.

[Del Bonifro et al. 2020] aimed to develop a model to predict freshman dropouts. Various approaches were explored to determine the most effective machine learning technique for their model: Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and Random Forest (RF). The results were categorized by schools based on the samples and division of students, yielding diverse yet significant findings for dropout prediction.

[Viloria et al. 2020] proposes a framework of early detection systems for potential dropouts by using four techniques: decision trees, logistic regression, Naive Bayes and K-Nearest Neighbors. The estimate of the project's capacity to predict university dropout is greater than 54%, and its precision is greater than 84%.

[Nagy and Molontay 2023] aimed to differentiate between students expected to graduate and those at risk of dropping out. Authors interpret their black-box machine learning model on a global and local scale. They first identify the most influential features and analyze their impact on the model's output. Additionally, they compare this to the inherently interpretable logistic regression. Furthermore, the authors assess the output and readability of various Explainable AI (XAI) tools.

[Santos et al. 2020] uses machine learning based on decision trees with genetic algorithms, researching to predict whether a student will evade or not. Their main objectives are to provide insights into potential students who may drop out to give improvements to the academic management of undergraduate programs and prevent at-risk students from evasion.

[Tenpipat and Akkarajitsakul 2020] focus on studying factors affecting undergraduates' educational status and creating binary classification models for predicting their situation, whether they will be dropouts or other statuses. The authors use data mining and machine learning techniques to develop three classifiers based on a decision tree, Random Forest, and Gradient Boosting classification. The results show that the prediction accuracy of gradient boosting, decision tree, and RF models are 93%, 92%, and 92%, respectively.

Our research provides an overview of a Computer Science Education program in a federal university from a city in a non-metropolitan area and its dropout rates. Our primary focus lies in analyzing a dataset specific to this program. This dataset was subjected to a wide array of machine learning algorithms. The objective was to identify, through the analysis of several variables, a subset of the most significant ones that contribute to identifying dropout cases.

3. Proposed approach

The addressed program is a Bachelor of Computer Science Education at the Federal University of Paraná (UFPR), which emphasizes developing educational and technical computer science skills. We will refer to it as the "CSE program."

Operating since 2014, the CSE program spans four years with evening classes, allowing students to work or engage in research. Situated in Palotina, a rural town in Paraná (Brazil) with about 30,000 residents, the program draws students mainly from smaller neighboring towns. Additionally, non-metropolitan regions tend to have higher dropout rates than large urban centers [Barbosa-Camargo et al. 2021, Lewine et al. 2021].

3.1. CSE program dropout scenario

To better understand the student dropout phenomenon within the CSE program, we collected and analyzed historical data provided by the UFPR's student management system, as depicted in Figure 1.

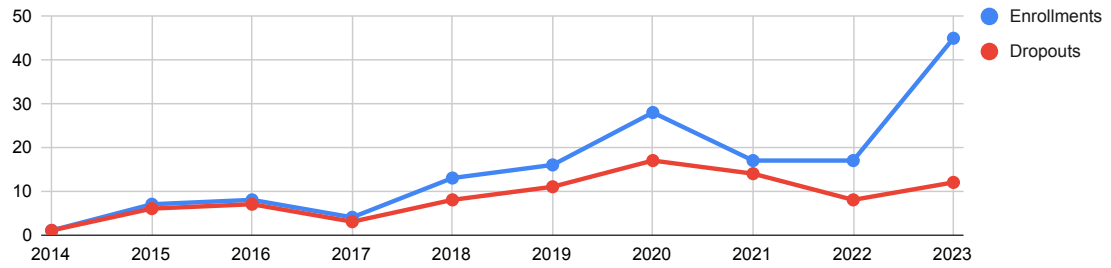

Figure 1. Students enrollments and dropouts by enrollment year.

The data reveals a striking pattern: many students who enrolled in the CSE program between 2014 and 2017 have ultimately discontinued their studies. While there is a slight improvement post-2018, dropouts still constitute a substantial proportion of the enrollments.

We then analyzed how long students stay in the CSE program, whether they complete it or choose to discontinue their studies (Table 1).

Table 1. Time spent within the CSE program in semesters.

	Mean	STD	MAX	Median	MIN
Conclusion	11.6	2.2	14	10	10
Dropout	4.8	3.9	17	4	0

The data is presented in semesters, indicating that students typically require approximately 11 semesters to graduate from the program. Concerning dropouts, they typically exit the program after about 5 semesters, with a notable standard deviation of approximately 4 semesters. However, the median value suggests that half of all dropouts occur within the first 5 semesters.

Given the temporal nature of our data analysis, assessing the program's completion rates over time is crucial to pinpoint when students are most vulnerable to discontinuing their studies. Figure 2 illustrates the mean percentage of program completion among dropout students across different semesters. In this graph, the left y-axis (blue bars) represents the number of dropout students, while the right y-axis (red line) represents the program completion rate based on the number of completed courses. In the fourth semester, for example, the average program completion rate is 3.37%, corresponding to about 2 completed courses.

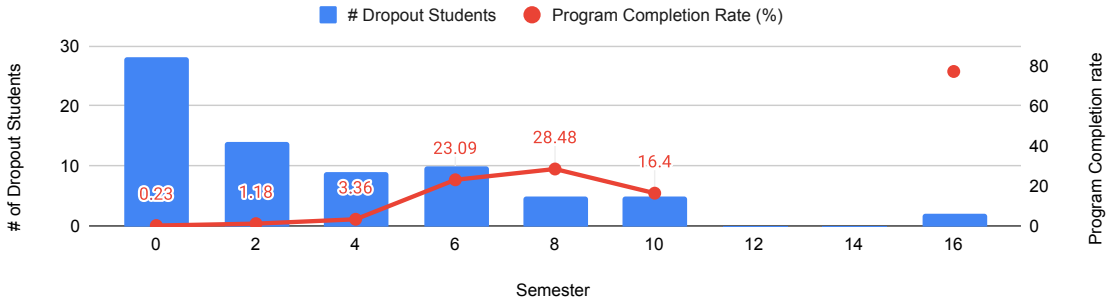

Figure 2. Number of dropout students versus program completion rate by semester.

As evident, most dropout incidents are concentrated within the first 4 semesters of the CSE program. Furthermore, during this initial period, the program completion rate for

these students averages less than 4%, significantly lower than the maximum completion rate of 30% observed across all dropout students.

To address the issue of dropouts comprehensively, we deemed it necessary to explore the root causes behind these occurrences.

Figure 3 provides an overview of the primary causes of dropouts in the CSE program.

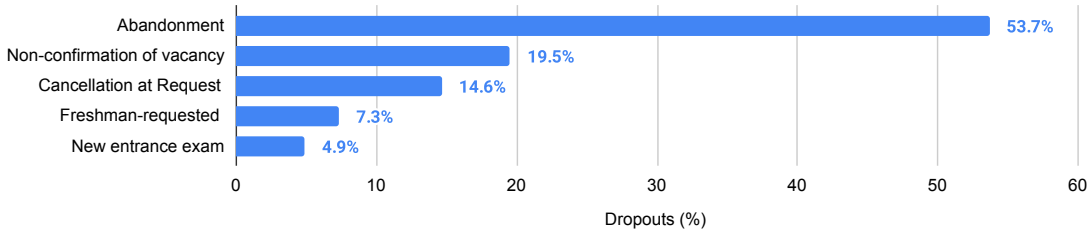

Figure 3. Causes of dropout in the CSE program.

The data illustrates that the most prevalent cause of dropout is abandonment, followed by non-confirmation of vacancy and cancellation at request. This finding aligns with our previous observations, suggesting that the high dropout rates within the program's initial 4 semesters may trigger a cascade effect. If underlying issues persist unnoticed during these early stages and students lack appropriate academic guidance, the resulting lack of progress and increasing reintegration difficulties can lead to students dropping out.

3.2. Dataset

Given the significant dropout rate within the initial 4 semesters, especially among students with low completion rates, our primary goal is to proactively identify potential dropouts based on their information.

We collected data from two UFPR datasets – one with socio-demographic details (26 features, including city, enrollment year, birth date, admission exam type, course failures, and academic performance) for all CSE students and another focused on individual course performance (attendance rates, final scores, codes, and names). Our model prioritizes privacy, excluding sensitive data like race or gender, ensuring a responsible and unbiased approach.

Among the preprocessing steps performed is the standardization of the target variable "situation": active and graduated students were considered non-dropouts (0), while course abandonments, regardless of the reason, were considered as dropouts (1). Additionally, first-year students were excluded from the database, as they had not yet completed all first-year courses. The final database contains 113 records used for training, testing, and validating the evaluated models.

It was crucial to include demographic characteristics in our analysis due to the complex interaction between socioeconomic factors in the CSE program, which is located in a small city. Many students from smaller neighboring towns face challenges such as transportation limitations, which impact their attendance and academic performance. Therefore, by considering demographics, we aim to address these barriers in dropout prevention.

We identified eight students facing academic challenges while enrolled using insights from EDA and internal reports. Recognizing them as potential dropouts, we removed them from the dataset but kept their data for validation testing.

3.3. ML models' implementation

With the dataset ready, we implemented the ML classifier using PyCaret [Ali 2020], which aided in automating ML training and comparison. For robust evaluation, 20% of the dataset was reserved for testing (details in Section 4). The rest underwent a 10-fold cross-validation by PyCaret, as shown in Figure 4.

```
from pycaret.classification import *
s = setup(data=train, target = 'situacao',
          session_id = 2, train_size=0.8,
          remove_multicollinearity=True,
          multicollinearity_threshold=0.8,
          max_encoding_ohe=15,
          )
```

Figure 4. Pycaret experiment setup.

PyCaret was configured with "remove_multicollinearity" true for improved performance and managing multicollinearity. This setting automatically eliminated highly correlated features with a "multicollinearity_threshold" of 0.8. We also set "max_encoding_ohe" to 15, limiting one-hot encoding to features with up to 15 categories, while high cardinality features used target encoding.

Our approach involved evaluating 15 distinct ML methods, each offering unique strengths and capabilities: Ada Boost, Decision Tree, Dummy Classifier, Extra Trees, XGboost, Gradient Boosting, kNN, LGB, Linear Discriminant Analysis, Logistic Regression, Naive Bayes, Quadratic Discriminant Analysis, Random Forest, Ridge Classifier, and SVM.

Anonymized datasets, as well as the preprocessing and implementation routines, are available on GitHub[2].

4. Results

As previously indicated, our approach was evaluated using real data from students enrolled in the CSE program at the Federal University of Paraná (UFPR), Brazil.

To enable an in-depth evaluation, we partitioned 20% of the dataset for performance testing, with the remaining portion utilized for model training. Additionally, the dataset underwent careful curation, which involved omitting 8 rows containing information about students grappling with academic challenges. These excluded rows were preserved for a final validation test (Section 4.2).

Then, we performed a feature selection process on the dataset, which contained 68 features about each student, considering the impracticality of using all this data for the advisory team. Feature importance scores were collected from methods supporting the "$coef\_$" or "$feature\_importances\_$" attributes, and the most critical and frequent features for each ML method were selected. "University time" and the students' city of

[2]https://github.com/mvoassis/dropout_prediction

origin were highlighted as highly significant features. The selected features are shown in Table 2, where features beginning with "DEE" represent course codes within the CSE program.

Table 2. Resulting most relevant features.

#	Feature	Course	Semester
1	University time	-	-
2	City	-	-
3	Year of School completion	-	-
4	Academic Performance Index	-	-
5	DEE345 (Score)	Computer Systems Security	2^{nd}
6	DEE341 (Attendance)	Programming Laboratory I	1^{st}
7	DEE374 (Attendance)	Pre-Calculus	1^{st}
8	Failures due to attendance	-	-

All the relevant courses related to student dropout belong to the technical courses of the CSE program. Thus, there is no direct strong correlation between early dropouts and courses from the educational/teaching part of the program.

4.1. Choosing the ML model

In this phase, we applied the PyCaret library to automate the evaluation of our proposed classifier. By default, PyCaret employs a 10-fold stratified cross-validation procedure for presenting the classification outcomes. Table 3 summarizes the classification results.

Table 3. Classification performance metrics (8 features).

Model	Accuracy	AUC	Recall	Precision	F1-Score
Logistic Regression	**0.9**	**0.955**	0.95	0.85	**0.8944**
Ada Boost Classifier	0.8889	0.91	0.875	**0.885**	0.8694
Extra Trees Classifier	0.8444	0.905	0.85	0.8567	0.8317
XGBoost	0.7556	0.905	0.695	0.775	0.6996
Random Forest Classifier	0.7444	0.9025	0.67	0.7895	0.6839
Linear Discriminant Analysis	0.7667	0.895	0.74	0.765	0.7279
Gradient Boosting Classifier	0.7333	0.885	0.72	0.7545	0.6873
Naive Bayes	0.6333	0.865	**1**	0.5554	0.7133
K Neighbors Classifier	0.7889	0.8425	0.825	0.7521	0.7811
LightGBM	0.7556	0.835	0.695	0.7967	0.6922
Quadratic Discriminant Analysis	0.7111	0.81	0.665	0.691	0.6315
Decision Tree Classifier	0.7111	0.71	0.62	0.7717	0.6333
Dummy Classifier	0.5444	0.5	0	0	0
SVM - Linear Kernel	0.5778	0	0.3	0.4873	0.3028
Ridge Classifier	0.8222	0	0.81	0.8	0.7885

The outcomes, sorted by AUC, indicate Logistic Regression as the top AUC, accuracy, and F1 score method. Ada Boost, Extra Trees, XGBoost, and Random Forest closely follow with $AUC > 0.9$.

Despite the results of Logistic Regression, the system's primary goal is to identify potential dropout students in the CSE Program. Thus, occasional misclassification of enrolled students with dropout-like attributes (false positives) is essential. Simultaneously, minimizing false negatives is crucial to avoid misclassifying actual dropout students as enrolled. To evaluate this, we collected FP and FN values from the test dataset classification, as shown in Figure 5.

The numeric values indicate the number of false positives per method. While Logistic Regression achieved the highest average results, it failed to identify any potential dropout students (false negatives) and even misclassified one dropout student as an enrolled one. In contrast, seven tested methods generated false-positive results, signifying their potential to identify at least one probable dropout student.

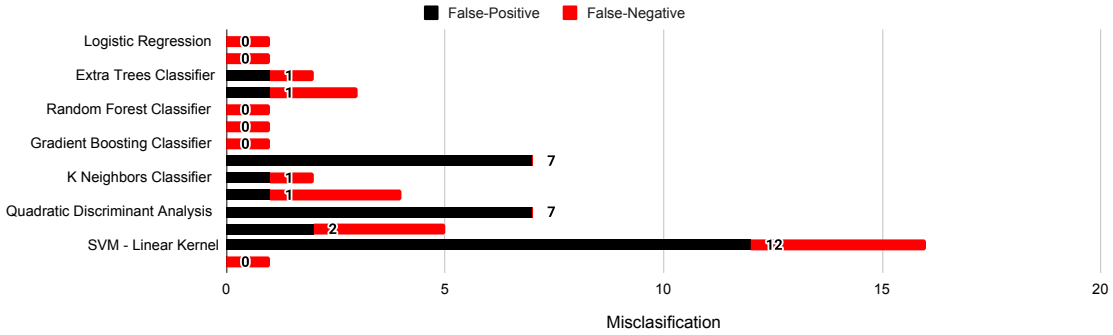

Figure 5. False-positives and false-negatives per method.

We meticulously assessed the information in these false-positive cases through consultations with the advisory team and the university system data. Following this thorough investigation, we concluded that three students from the test dataset could be considered potential dropouts. Consequently, we eliminated using Naive Bayes and Quadratic Discriminant Analysis, as both models misclassified four regularly enrolled students.

Subsequently, we analyzed the five remaining classification methods: Extra Trees, XGBoost, kNN, LightGBM, and Decision Tree.

We subjected them to hyperparameter optimization to further refine the model selection process, which was also facilitated through PyCaret. The models underwent optimization with Recall as the primary guiding metric to improve their performance concerning the classification of positive cases. The hyperparameter optimization was conducted through 50 iterations of 10-fold cross-validation, totaling 500 fits. The classification is depicted in Table 4.

Table 4. Classification performance metrics (hyperparameter optimization).

Model	Accuracy	AUC	Recall	Precision	F1-Score
LightGBM	0.8222	**0.905**	0.835	0.7967	0.8044
Extra Trees Classifier	0.8222	0.89	0.85	0.7967	0.8127
XGBoost	0.7444	0.86	**1**	0.6586	0.7865
K Neighbors Classifier	**0.8556**	0.8525	0.825	**0.8467**	**0.8317**
Decision Tree Classifier	0.8111	0.8325	0.875	0.7483	0.796

Regarding classification metrics, LightGBM excelled in AUC, kNN demonstrated better Accuracy, Precision, and F1 metrics, while XGBoost achieved the highest Recall value. However, as discussed earlier, we must incorporate the analysis of false-positive and false-negative results to support our model choice, as shown in Figure 6.

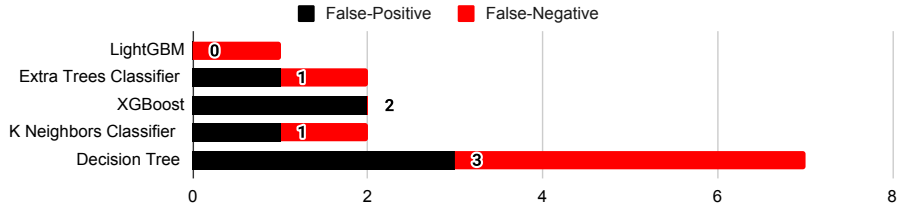

Figure 6. False-positives and false-negatives per method after hyperparameter optimization.

LightGBM generated no false-positive values, indicating its inability to identify potential dropout students. Decision Tree and XGBoost performed admirably in terms of identifying possible dropout students. Decision Tree successfully identified the three

potential dropout students in the test dataset but misclassified four actual dropout students as enrolled students. Therefore, after careful consideration, we selected the XGBoost Classifier as the primary engine for our dropout detection system. The optimized hyperparameter values are available on GitHub[3].

Table 5 provides the 10-fold cross-validation results for the tuned XGBoost, presenting both mean results and their standard deviation.

Table 5. Tuned XGBoost cross-validation outcomes.

Fold	Accuracy	AUC	Recall	Precision	F1-Score
0	1	1	1	1	1
1	0.6667	1	1	0.5714	0.7273
2	0.7778	0.8	1	0.6667	0.8
3	0.7778	0.9	1	0.6667	0.8
4	0.5556	0.65	1	0.5	0.6667
5	0.8889	1	1	0.8	0.8889
6	0.6667	0.65	1	0.5714	0.7273
7	0.7778	0.95	1	0.6667	0.8
8	0.6667	0.8	1	0.5714	0.7273
9	0.6667	0.85	1	0.5714	0.7273
Mean	0.7444	0.86	1	0.6586	0.7865
Std	0.1222	0.1281	0	0.1387	0.0922

XGBoost achieved the maximum Recall score for all folds, indicating its ability to successfully identify all dropout cases. Moreover, the high standard deviation observed in Precision, AUC, and F1 metrics suggests an imbalance in classification scores, implying that the model may generate some false-positive values, which aligns with our system's objective.

4.2. Validation with Known Potential Dropout Students

As detailed in Section 3.3, we intentionally excluded eight rows from the dataset used for training and testing. These rows represent students with documented academic challenges and are considered potential dropouts by our advisory team. These students are presently enrolled in the CSE program, so their classification as dropouts would essentially amount to false positives, serving as a crucial metric to gauge the system's effectiveness in addressing the specified problem.

The anonymized data of these specific students is presented in Table 6.

Table 6. Potential Dropout Students' data.

ID	U.Time	City	HS comp.	API	DEE345	DEE341	DEE374	Fail. (A)
1	8	Assis Chat.	2019	0.2303	0	40	25	21
2	8	Palotina	0	0.1613	0	15	31.4	17
3	10	Assis Chat.	0	0.1321	0	23.5	38.75	25
4	18	Palotina	2011	0.41	12	0	0	18
5	8	Palotina	2018	0.2209	0	74	48.5	16
6	8	Curitiba	2018	0.1469	0	53.5	25	20
7	8	Palotina	0	0.1386	0	15	31.4	21
8	12	Me. Rondon	2012	0.3519	0	0	88	24

Subsequently, we subjected this data to the tuned XGBoost model, and the resulting confusion matrix is illustrated in Figure 7.

Figure 7 shows that the optimized XGBoost effectively identified potential dropouts among active CSE program students. It highlights the model's proficiency in

[3]https://github.com/mvoassis/dropout_prediction

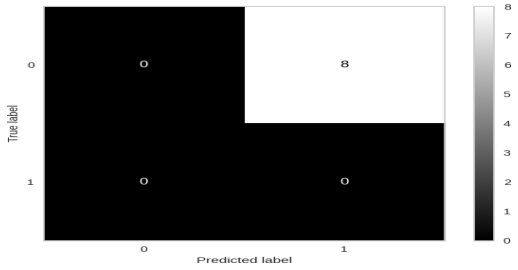

Figure 7. Confusion matrix - known potential dropout students.

recognizing high-risk cases and offers insights into our system's effectiveness in identifying students at risk of discontinuing studies.

4.3. Discussion on Feature Relevance and Mitigation Actions

Although the proposed model's results in detecting potential dropouts are promising, the selected features offer vital insights into understanding and mitigating causes. We discuss each feature's relevance and potential implications for pedagogical interventions and institutional policies. We draw from our tacit knowledge of over ten years of experience, including three years as program heads, for a qualitative evaluation of this specific program.

University Time: This feature reflects the time elapsed since the student enrolled. Students who have been in the program for longer may face challenges such as demotivation, lack of identification with the field, or even financial difficulties, which may lead to dropout. Although 50% of dropouts occur within the first four semesters of the program, the increase in time spent in the program correlates with a higher probability of dropout, reinforcing the need for early interventions to mitigate this issue.

City: The hometown plays a significant role in students' adaptation to the university environment, especially where most students come from smaller, nearby cities. It is observed that the dropout rate among students from Palotina is 41%. In contrast, in cities like Maripá, Iporã, Assis Chateaubriand, and Marechal Cândido Rondon, the dropout rates are significantly higher, being 70%, 67%, 57%, and 56%, respectively. For students from these cities, where transportation to Palotina is exclusively private or, in the case of Assis Chateaubriand, non-existent, the situation is even more challenging. These students rely on their means for commuting, which can result in higher costs, less time available for studies, and a sense of isolation from academic life.

Year of School Completion: The year of completion of basic education may indicate the time gap between the end of high school and university enrollment. A larger gap may be associated with difficulties in readjusting to the study environment and academic content.

Academic Performance Index: It is an institutional metric that reflects the student's performance in their courses, including grades and attendance. As expected, this index proved to be a strong predictor of dropout, where students with lower indices are more likely to abandon the course.

Score in the Course "Computer Systems Security": This course is offered in a distance learning format, which may present an additional challenge, especially for students who are at the beginning of higher education and have not yet fully developed

the autonomy required for independent study. When entering university, many students report difficulties managing their time and study resources without the structure of in-person classes. It may explain why low grades in this course are associated with a higher dropout risk.

Attendance in the Courses "Programming Laboratory I" and "Pre-calculus": These courses cover introductory concepts for the program's students. The "Programming Laboratory I" course introduces basic programming concepts, while "Pre-calculus" reviews high school math content. Low attendance in both courses was observed to be strongly correlated with a higher risk of dropout.

One factor that may contribute to this difficulty is the low competition in the entrance exam, which can result in the selection of candidates who arrive at university with significant gaps in their basic education. When faced with fundamental content in programming and mathematics, these students may face considerable challenges due to the complexity of the subjects and the lack of adequate prior preparation. This combination of factors can lead to demotivation and, eventually, dropout.

Failures due to Attendance: The number of failures due to attendance directly indicates engagement problems and possibly external issues, such as work or personal matters that prevent regular class attendance.

The analysis of these features indicates clear areas where specific interventions can be implemented to reduce dropout rates. Some measures can be implemented to mitigate the dropouts, starting with strengthening academic and emotional support for students from the early semesters. Tutoring and reinforcement programs in initial courses can help students overcome knowledge gaps and adaptation difficulties. Finally, creating continuous monitoring mechanisms for academic performance, such as the approach proposed in this work, and early intervention for students with low academic performance or irregular attendance can significantly reduce dropout rates.

5. Conclusion

Our study utilized machine learning to predict student dropout risk in a Computer Science Education (CSE) program, employing a dataset from the Federal University of Paraná (UFPR), Brazil. Among 15 algorithms tested, XGBoost demonstrated the highest performance (recall=1), effectively identifying all dropout cases in the test and validation sets.

We identified critical predictors of dropout risk, including total time spent in the university program, city of origin, high school completion year, academic performance index, and score/attendance in technical courses. Early intervention, especially within the first 4 semesters, could be highly impactful.

Our tuned XGBoost model effectively identified potential dropouts, emphasizing the value of AI in enhancing student retention. This proactive approach, offering targeted support, can significantly improve graduation rates.

Future research could incorporate additional features related to student demographics and qualitative feedback to enhance predictive capabilities further.

References

Ali, M. (2020). *PyCaret: An open source, low-code machine learning library in Python*. PyCaret version 3.0.

Barbosa-Camargo, M. I., García-Sánchez, A., and Ridao-Carlini, M. L. (2021). Inequality and dropout in higher education in colombia. a multilevel analysis of regional differences, institutions, and field of study. *Mathematics*, 9(24).

BRASIL (2022). Parecer cne/cebnº2/2022: Normas sobre computação na educação básica – complemento à base nacional comum curricular (bncc) (*only in portuguese*). Technical report. Available `http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=235511-pceb002-22&category_slug=fevereiro-2022-pdf&Itemid=30192`. Accessed: 09 Oct. 2023.

Bravo, D., Alves, M. Z., Ensina, L., and de Oliveira, L. (2023). Evaluating strategies to predict student dropout of a bachelor's degree in computer science. In *Anais do XI Symposium on Knowledge Discovery, Mining and Learning*, pages 1–8, Porto Alegre, RS, Brasil. SBC.

Colpo, M., Primo, T., and Aguiar, M. (2021). Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 873–884, Porto Alegre, RS, Brasil. SBC.

de Almeida, C. C. and Mateus, N. M. A. (2015). Licenciandos em computação: experiências formativas proporcionadas pelo pibid e a busca pelo reconhecimento profissional. *Horizontes*, 33(1).

Del Bonifro, F., Gabbrielli, M., Lisanti, G., and Zingaro, S. P. (2020). Student dropout prediction. In Bittencourt, I. I., Cukurova, M., Muldner, K., Luckin, R., and Millán, E., editors, *Artificial Intelligence in Education*, pages 129–140, Cham. Springer.

Fluminense, U. F. (2015). Forplad - indicadores. In *Fórum de Pró-Reitores de Planejamento e Administração Comissão de Planejamento e Avaliação*. Available `https://www.uff.br/sites/default/files/indicadores_do_forplad.pdf`.

Lewine, R., Manley, K., Bailey, G., Warnecke, A., Davis, D., and Sommers, A. (2021). College success among students from disadvantaged backgrounds: "poor" and "rural" do not spell failure. *Journal of College Student Retention: Research, Theory & Practice*, 23(3):686–698.

Linhares, A. C. O. and Santos, K. S. (2021). A licenciatura em computação no brasil: histórica e contexto atual. *Revista Brasileira de Informática na Educação*, 29:188–208.

Mathews de, N. S. L., Fachini Gomes, J. B., Holanda, M., Koike, C. C., and Leao Costa, M. T. (2023). Study on computer science undergraduate students dropout at the university of brasilia. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–7.

Moscoviz, L., Evans, D. K., et al. (2022). Learning loss and student dropouts during the covid-19 pandemic: A review of the evidence two years after schools shut down. *Center for Global Development Washington, DC, USA*.

Nagy, M. and Molontay, R. (2023). Interpretable dropout prediction: Towards xai-based personalized intervention. *Int. J. of Artificial Intelligence in Educ.*, pages 1–27.

Nascimento, P. A. M. M. and Verhine, R. E. (2017). Considerações sobre o investimento público em educação superior no brasil. Available `https://repositorio.ipea.gov.br/handle/11058/7648`. Accessed: 08 Mar. 2024.

Olmedo-Cifuentes, I. and Martínez-León, I. M. (2022). University dropout intention: Analysis during covid-19. *Journal of Management and Business Education*, 5(2).

Santos, G., Belloze, K. T., Tarrataca, L., Haddad, D. B., Bordignon, A. L., and Brandão, D. N. (2020). Evolvedtree: Analyzing student dropout in universities. In *2020 Int. Conf. on Systems, Signals and Image Processing (IWSSIP)*, pages 173–178. IEEE.

Santos, J., Sousa, J. D., Mello, R., Cristino, C., and Alves, G. (2021). Um modelo para análise do impacto da retenção e evasão no ensino superior utilizando cadeias de markov absorventes. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 813–823, Porto Alegre, RS, Brasil. SBC.

Schefer-Wenzl, S., Miladinovic, I., Bachinger-Raithofer, S., and Muckenhumer, C. (2024). A study on reasons for student dropouts in a computer science bachelor's degree program. In Auer, M. E., Cukierman, U. R., Vendrell Vidal, E., and Tovar Caro, E., editors, *Towards a Hybrid, Flexible and Socially Engaged Higher Education*, pages 391–400, Cham. Springer Nature Switzerland.

Tenpipat, W. and Akkarajitsakul, K. (2020). Student dropout prediction: A kmutt case study. In *2020 1st Int. Conf. on Big Data Analytics and Practices*, pages 1–5.

Varga, E. B. and Ádám Sátán (2021). Detecting at-risk students on computer science bachelor programs based on pre-enrollment characteristics. *Hungarian Educational Research Journal*, 11(3):297 – 310.

Viloria, A., Naveda, A. S., Palma, H. H., Núñez, W. N., and Núñez, L. N. (2020). Using big data to determine potential dropouts in higher education. *Journal of Physics*.

Wegner, R. C. (2022). Evasão no ensino superior: Digressões motivadas a partir da pandemia do novo coronavírus. *Revista Docência e Cibercultura*, 6(1):01–22.