Evaluating Strategies to Predict Student Dropout of a Bachelor's Degree in Computer Science

Dayane Perez Bravo[1], Marco Antonio Zanata Alves[1], Leandro Augusto Ensina[1,2], Luiz Eduardo Soares de Oliveira[1]

[1] Federal University of Paraná (UFPR), Brazil
dayaneperezbravo@gmail.com, mazalves@inf.ufpr.br, lesoliveira@inf.ufpr.br
[2] Federal University of Technology – Paraná (UTFPR), Brazil
leandroa@utfpr.edu.br

Abstract. The Brazilian Higher Education Census has revealed that the dropout rate among higher education students in Brazil exceeds 50% starting from the fifth year. This high rate results in several problems related to the wastage of resources invested by both the society and the students. Therefore, universities need to develop strategies to prevent student dropout and minimize these problems. However, predicting student dropout involves detecting patterns and predicting them over a large amount of data collected yearly from thousands of students. Given the scale and volume of data involved in dropout prediction, machine learning emerges as a powerful technique to automate the identification of these students. The objective of this paper is to identify students who are prone to dropping out based on the academic history of Bachelor's Degree in Computer Science students at an unpaid public university in Brazil. We engineered four datasets based on the semester in which the students are enrolled. These datasets are designed to simulate the academic scenario and individual characteristics of the students available up to the prediction moment. Besides, we propose three feature models to identify the best scenario. Our method could identify the students most likely to drop out and the main features that contributed to the respective decision. Using only the information from the disciplines taken by the students proved to be the best feature model. When using these features with Gradient-Boosting, the F1-Score performance ranged between 69% and 85%, depending on the dataset.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: data mining, feature engineering, pattern recognition, student evasion

1. INTRODUCTION

The Brazilian Institute of Research on Education (Inep) [INEP 2022] conducted the Brazilian Higher Education Census between 2011 and 2020 to examine the retention rates of students enrolled in undergraduate programs in Brazil. The findings from the initial five years of monitoring revealed a dropout rate of 51% among the entrants, with only 29% successfully completing their courses. Subsequently, after ten years of follow-up, 40% of the initial group managed to complete their programs, while 59% dropped out. These statistics highlight that in Brazil, where undergraduate courses typically last for five years, more than half of the students discontinue their studies within this timeframe. This emphasizes the need to establish effective student retention policies to tackle this issue.

By implementing effective student retention policies in higher education, the negative consequences of dropouts can be mitigated. It is therefore essential to comprehend the underlying reasons for dropout and identify students who are at a higher risk of discontinuing their courses. These predictions can aid Higher Education Institutions (HEI) in developing proactive measures to encourage students to persist in their studies. However, such predictions often involve dealing with a substantial volume of data, given the scale of the problem. Consequently, the utilization of Machine Learning (ML)

This work was partially supported by Serrapilheira Institute (grant number Serra-1709-16621).

algorithms is highly recommended to automate the identification of these students [Alban and Mauricio 2019; Santos et al. 2018].

This paper aims to employ ML algorithms to identify students who are prone to dropout. Specifically, we focus on analyzing the academic history of students enrolled in a Bachelor's Degree in Computer Science (BCC) program at the Federal University of Parana (UFPR), an unpaid public university in Brazil. We organized the original database into four distinct datasets corresponding to different semesters of the course (3rd, 5th, 7th, and 9th). These datasets allow us to simulate various scenarios that students encounter throughout their academic journey, while considering only the available information up until the prediction point. Furthermore, we developed three feature models to determine the important attributes that contribute to accurate predictions of students at a higher risk of dropping out.

The significant contributions of this research are as follows: (i) the implementation of feature engineering techniques to prepare the datasets, enabling the prediction of student dropout at different stages of the course; (ii) the identification of students who are at an elevated risk of dropping out; and (iii) the development of a model tailored specifically to the BCC course offered by a public and unpaid university in Brazil.

This paper is organized as follows. Section 2 describes some related work. Section 3 outlines our method by the data understanding, the evaluated models, and the experimental protocol. Section 4 reports and discusses our results. Finally, Section 5 highlights the strengths and limitations of the proposed method, including suggestions for future research.

2. RELATED WORK

[Romero and Ventura 2020] have highlighted the significance of Educational Data Mining (EDM) in supporting HEI by utilizing available data, ML techniques, and an understanding of the educational management system. They noted that previous studies conducted between 2000 and 2018 have applied EDM to predict student dropout. [Alban and Mauricio 2019] conducted a survey revealing that classifiers based on decision trees constituted 79% of the research on student dropout prediction published between 2006 and 2018. In alignment with these findings, our proposal involves employing ML classifiers based on decision trees to predict students at risk of dropout.

[Fernández-García et al. 2021] conducted a study utilizing personal, demographic, and academic data to predict students who are most likely to dropout. They evaluated the performance of Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF) classifiers using five different models. The first model focused on the moment of enrollment, while the other models were based on the first four semesters of the students' academic journey. They reported an accuracy ranging from 70% to 91%, with the lowest score observed in the first model and the highest in the fifth model. In our research, we were unable to replicate their work due to the unavailability of certain features used in their study, such as address and birth information. Furthermore, our datasets are based on different time periods, the first four years rather than just the first four semesters. Additionally, we include four additional classifiers: Decision Tree (DT), Extra-Trees (ET), AdaBoost (AB) and Logistic Regression (LR). As a result, we compare the performance of a total of seven classifiers in our study.

[Santos et al. 2021] conducted a study in which they utilized DT and RF classifiers to predict student dropout. They achieved accuracy rates ranging from 79.31% to 98.25% using solely the academic performance data of students across each semester of their course. The authors assert that their models can be replicated by others HEI. In a similar vein, we propose three attribute models to evaluate the effectiveness of a model based solely on academic data. The first model includes only the personal features of the students, the second model includes only the academic features, and the third model incorporates both personal and academic features. It is important to note that our datasets are structured based on years rather than semesters, and we employ the seven classifiers previously mentioned to assess the predictive performance.

3. PROPOSED METHOD

In this study, we utilized a database comprising anonymous information from 2,763 students who enrolled in the BCC course at UFPR between 1995 and 2019. The personal data of the students included in our database encompass the following attributes: student registry code, gender, year of the curriculum, forms of admission and evasion, as well as the year and period in which admission and evasion took place. Additionally, the database contains information regarding the disciplines undertaken by each student, including the code, name, final grade, year and period of attendance, and approval status.

3.1 Data Preparation

Throughout the BCC's history, several curricula were adopted according to the university's and students' best interests. The course made the most recent curriculum changes in 2011 and 2019. For our study, we specifically focused on students enrolled in the 2011 curriculum to ensure uniformity in terms of mandatory disciplines. These disciplines and the respective periods in which they were offered are available in reference [UFPR 2011]. The reference [UFPR 2011] provides detailed information about these disciplines and the corresponding periods in which they were offered. It is crucial to highlight that we excluded students who remained active in the course because their status regarding dropout or graduation is uncertain. Consequently, these students are not suitable for model training or testing since they lack the target value necessary for supervised learning.

To mitigate class imbalance issues, we carefully selected the years of admission with a balanced proportion of graduates and dropouts. As a result, we chose the years between 2006 and 2014 for our study. It is worth noting that students often transition to newer curricula as soon as they become available. This phenomenon explains why some students who entered before 2011 are enrolled in a more recent curriculum, further justifying our inclusion of students with an entry year prior to 2011. Also, we excluded disciplines that are not part of the regular course curriculum. Consequently, our final database consisted of 594 students, out of which 326 graduated, while 268 dropped out.

After thoroughly cleaning the database, we eliminated redundant data, special characters, and highly correlated attributes. The following attributes were removed from the dataset: student registry code, course name, course code, curriculum code, name of disciplines, and theoretical and practical hours of the disciplines. Using the available information, we generated new academic attributes for each student: (i) The semester of the course in which the student took each discipline. (ii) The semester in which the student dropped out, if applicable. (iii) The number of times the student took each discipline. (iv) Whether the student was on track with the regular curriculum schedule.

In our analysis, we defined a student as "periodized" if they successfully completed all their disciplines in the same semester as specified in the regular curriculum or if they completed them ahead of schedule. A student is considered "not periodized" if they took a discipline at a later semester than prescribed or if there was repetition, such as taking a discipline for the first time in the regular semester and then again in a subsequent semester. Furthermore, we observed that the personal data available for the students was limited. So, we incorporated the candidate/vacancy ratio provided by the UFPR agency responsible for the entrance selection exam [UFPR 2022]. This ratio, specific to each year of entry, provides additional contextual information about the competitiveness of the admission process for the students in our dataset.

Therefore, we have divided the available data into two distinct datasets: the **personal** dataset and the **academic** dataset. The **personal** dataset comprises the following features: gender, admission form, and candidate/vacancy ratio. On the other hand, the **academic** dataset includes the following features: discipline code, final grade, whether the discipline belongs to the regular curriculum, the semester of the course in which each discipline was taken, the number of times the student took each discipline, and whether the student is periodized with respect to the regular curriculum. It is important to note that the target variable of our study is the dropout status, which indicates whether a student dropped out or not.

3.2 Data Manipulation

We organized the data in a structured table format, with each student's attributes represented in a row. Academic data for disciplines were arranged sequentially. To handle cases where a student took a discipline multiple times, we opted to include only the data corresponding to the most recent instance of taking the discipline. This approach allows us to focus on the latest information available for analysis and prediction. By organizing the data in this manner, we can efficiently process and analyze the relevant features to develop our predictive models for identifying dropout-prone students.

We created four distinct datasets, A, B, C, and D, to simulate different stages of the course based on available data. **Dataset A** specifically represents students who have completed their first year and attended two semesters. The focus of this dataset is to predict dropout cases starting from the third semester onwards. To ensure consistency, only students whose dropout semester was equal to or higher than 3 were included in Dataset A. Students who had already dropped out were excluded from this dataset. Similarly, **Dataset B** represents the scenario of students who have completed their first two years and have attended four semesters. It is used to predict dropout from the fifth semester onwards. Only students whose dropout semester was equal to or higher than 5 were included in Dataset B. **Dataset C** simulates the scenario of students who have completed their first three years and have attended six semesters. It is used to predict dropout from the seventh semester onwards. Only students whose dropout semester was equal to or higher than 7 were included in Dataset C. Finally, **Dataset D** represents the scenario of students who have completed their entire course and have attended eight semesters. It is used to predict dropout from the ninth semester onwards. Only students whose dropout semester was equal to or higher than 9 were included in Dataset D. By creating these datasets, we can analyze the dropout patterns at different stages of the course and develop targeted predictive models for each scenario.

For Dataset A, we included only data from mandatory disciplines with semesters lower than 3. This was done because the number of students with anticipated disciplines was small (less than 3%), and including data from other disciplines could introduce bias. In the same way, for Datasets B, C, and D we selected data from mandatory disciplines with semesters lower than 5, 7, and 9, respectively.

3.2.1 *Data Splitting.* To ensure consistency in the rates of dropouts and graduates, we carefully analyzed each of the four generated datasets. The rates of dropouts and graduates from the initial datasets were maintained in the training and testing datasets. In the training phase, we randomly selected 80% of the students for training and the remaining 20% for testing. Throughout the model development, we exclusively used the training data. Within the training dataset, we further divided it by randomly selecting 20% of the students for the validation dataset and the remaining 80% for training. Finally, we utilized the testing datasets to validate and report the results of our models.

3.3 Models and Evaluation Methods

We developed three feature models for this research. **Model 1** includes only the three available personal features of each student: gender, admission form, and candidate/vacancy ratio. It consists of a fixed number of three attributes, regardless of the dataset. **Model 2** focuses on the disciplines taken by the students and includes four attributes for each discipline: final grade, semester within the course when the discipline was taken, the number of times the discipline was taken, and whether it was periodized concerning the regular schedule. The number of attributes varies for each dataset. For

Dataset A, which has 10 mandatory disciplines, there are 40 attributes. For Datasets B, C, and D, with 20, 30, and 34 mandatory disciplines, respectively, there are 80, 120, and 136 attributes. **Model 3** combines the features from both Model 1 and Model 2. It includes the three personal features for each student and the four discipline-related attributes for each discipline. The total number of attributes in Model 3 varies for each dataset. For Dataset A, it has 43 attributes, for Dataset B, C, with 83, 123, and 139 attributes.

Supervised learning algorithms were selected for this research as the data contains a labeled target variable with two possible classes: graduated or dropout. Hence, the task at hand is a classification problem. We considered seven classifiers in this study: (i) Decision Tree (DT) [Breiman et al. 1984]; (ii) Extra-Trees (ET), with 100 estimators [Geurts et al. 2006]; (iii) Random Forest (RF), with 100 estimators [Breiman 2001]; (iv) Gradient Boosting (GB), with 100 estimators [Friedman 2001]; (v) AdaBoost (AB), with 50 estimators [Zhu et al. 2009]; (vi) Support Vector Machine (SVM) [Chapelle et al. 2002]; (vii) Logistic Regression (LR) [Bishop 2007]. The chosen algorithms, including decision tree based algorithms, have been widely utilized in the literature for student dropout classification due to their effectiveness and accuracy [Alban and Mauricio 2019]. Additionally, considering the class imbalance in our dataset, with a difference in the number of graduates and dropouts, we chose to evaluate the classifiers using the Area Under the Curve (AUC) and F1-Score [He and Ma 2013].

4. RESULTS AND DISCUSSION

We assessed the performance of seven classifiers across three models and four datasets. Some original features were removed during validation as they did not enhance algorithm performance. After selecting the remaining discipline features, including code, final grade, and semester, we retrained the models and obtained the AUC performances, as shown in Table I.

Table I. Classifiers performances: AUC (%).

	Dataset	**AB**	**RF**	**GB**	**LR**	**SVM**	**ET**	**DT**
Model 1	A	61	61	64	56	59	64	60
	B	56	56	61	53	53	61	60
	C	55	55	59	53	50	59	56
	D	61	61	64	57	49	64	58
Model 2	A	65	65	67	73	68	52	58
	B	75	74	71	69	75	61	62
	C	76	87	85	76	83	76	82
	D	89	91	91	84	90	75	82
Model 3	A	60	60	56	62	68	47	58
	B	74	74	71	69	75	63	62
	C	79	84	84	78	83	71	82
	D	88	89	91	84	90	73	82

We also performed statistical tests to verify differences between models and classifiers. For all statistical tests, we used a significance level of 5%. Through Friedman's test, considering AUC results, we found that: (i) the classifiers have no statistical difference; and (ii) the models are statistically different from each other. Through the rankings analysis, we observed that: (a) Model 1 is significantly different from Models 2 and 3; and (b) Models 2 and 3 are statistically equivalents.

Therefore, we can infer that the statistical differences show that Models 2 and 3 are better than Model 1. This can be related to the fact that Models 2 and 3 have more relevant features than Model 1. On the other hand, Models 2 and 3 have no statistical difference, so we considered that Model 2 is better since it requires fewer attributes than Model 3.

With the selection of Model 2, we repeated the training step three times to re-test the statistical difference among the classifiers. Through Friedman's test, we can infer that the RF, GB, and SVM algorithms had the best performances. Among these three classifiers, only the RF and the GB allow identifying the attributes with the greatest influence (contribution/importance) on the predictions, an aspect that SVM does not allow. Consequently, we can use this aspect as a criterion to remove SVM and keep with RF and GB algorithms. We chose to present the GB results because we intend to evaluate the XGBoost algorithm [Chen and Guestrin 2016], which is based on GB, in future works. So, we can use this paper as a reference for this comparison. Table II shows the GB performances for each dataset considering precision, recall, F1-Score and AUC measures.

Table II. Performance of the proposed method for each dataset using GB.

	Precision	**Recall**	**F1-Score**	**AUC**
Dataset A	65%	72%	69%	67%
Dataset B	75%	74%	74%	71%
Dataset C	83%	77%	80%	85%
Dataset D	91%	80%	85%	91%

We used two methods supported by the GB algorithm to reinforce our analyzes: feature importance and predict probability. In particular, **feature importance** is a technique that calculates a score for each feature for a given model, in which the higher the value, the more relevant the feature. The relevance of a feature is computed based on the Gini importance [Ishwaran 2015]. So, we used this information to identify which features were more significant to the model's performance. In turn, **predicting probability** is a technique that calculates the probabilities of a given example belonging to each class (in our work, graduated or dropout). We used this measure to identify the probability of each student dropping out of the course.

Table II shows the GB performances for each dataset considering precision, recall, and F1-Score measures. Looking at the Precision column, we can see that the best performing model occurs on Dataset D, which is capable of identifying 91% of students who will drop out. We analyzed these results and observed an improvement in the AUC and F1-Score performances as we added information for more semesters. We expected this improvement in the performance as more semesters are used since more information about students is provided to the algorithms and, consequently, greater discriminative power for classification. This behavior explains the lower performance for Dataset A and the higher performance for Dataset D.

Analyzing the performance of GB in Dataset A concerning the confusion matrix, there are 71 successes and 43 errors. Most predictions performed correctly occurred with probabilities between 68% and 90%, approximately. For this dataset, the three most important features were the final grades of the disciplines: (i) "*Introduction to Algebra*", with 24% of contribution; (ii) "*Algorithms and Data Structures I*", with 17%; and (iii) "*Analytical Geometry*", with 15% of contribution to the model. These three disciplines are offered in the first semester of the regular curriculum.

The next analyzed performance of GB is in Dataset B, where there are 75 successes and 33 errors through the confusion matrix. Most predictions performed correctly occurred with probabilities between 90% and 93%. For this dataset, the three most important features were the final grades of the disciplines: (i) "*Digital Projects and Microprocessors*", with 30% of contribution; (ii) "*Introduction to Algebra*", with 9%; and (iii) "*Algorithms and Data Structures I*", with 7% of contribution to the model. These three disciplines are offered in the first semester too.

Now, the performance of GB in Dataset C shows 75 successes and 25 errors through the confusion matrix. Most predictions performed correctly occurred with probabilities between 90% and 100%. For this dataset, the three most important features were the final grades of the disciplines: (i) "*Basic Software I*", with 19% of contribution; (ii) "*Discrete Mathematics*", with 15%; and (iii) "*Computer Organization and Architecture*", with 7% of contribution to the model. These three disciplines are offered in the third semester.

Finally, the performance of GB in Dataset D shows 75 successes and 18 errors through the confusion matrix. Most predictions performed correctly occurred with probabilities between 90% and 100%. For this dataset, the three most important features were the final grades of the disciplines: (i) "*Discrete Mathematics*", with 36% of contribution to the model and offered in the third semester; (ii) "*Operational Systems*", with 8% and offered in the fourth semester; and (iii) "*Differential and Integral Calculus II*", with 6% of contribution to the model and offered in the third semester.

In addition, predictions performed correctly began accumulating at higher predict probabilities, while predictions performed amiss started to reduce in number at high probabilities. In this way, we can identify students prone or not to drop out with greater confidence. Finally, for all datasets, the features that present importance higher than 5% correspond to the attributes of the final grades of the disciplines. While features like "*Semester in which the discipline was taken*" demonstrate relevance lower than 5%. In addition, if we know the disciplines with the greatest contributions to the model, some actions can be taken, such as reinforcement classes and monitoring. We noticed that the first three semesters' disciplines had the highest contribution rates to the model. So, we can suggest that educational managers take some actions to identify the main problems related to these disciplines.

In a real-world scenario, datasets A, B, C, and D can be used simultaneously for different periods of the course. In this aspect, educational managers must select the students enrolled in the respective semesters of each dataset. With the dropout-prone students in hand, education managers can contact these students and take personalized actions on a case-by-case basis. These personalized actions can be applied even to students who tend to stay in the course but with lower probability (percentage obtained through predict probability function) and could become future dropouts. Another suggestion would be to refer these students to the pedagogy sector and, if applicable, to the institutional psychologist for more targeted guidance. These actions could be interesting to collect directly from students the reasons that led them to drop out.

5. CONCLUSION

In this paper, we proposed a ML method to predict student dropout in a BCC course at UFPR, a Brazilian public university. We created four datasets (A, B, C, and D) to simulate different stages of the course, corresponding to the beginning of semesters 3, 5, 7, and 9, respectively. This allowed us to evaluate multiple algorithms and three feature models.

We found that Models 2 and 3 outperformed Model 1. Model 2 was considered the best since it requires less data. The most influential features were the final grades of the disciplines. Although our method was developed specifically for the BCC course, it can be adapted for other courses in exact, biological, and human sciences in both public and private institutions. This adaptability is possible because our best model (Model 2) uses only information from the regular curriculum's disciplines specific to the chosen undergraduate course. Our method can be tailored to different undergraduate courses by incorporating their respective discipline data, as [Santos et al. 2021] suggested.

Although Model 1 had the worst performance, in future work, we recommend including additional personal data such as address, marital status, employment status, family income, living arrangements, and number of children to improve the performance of the dropout prediction model. A model with more personal data is likely to perform better. Our hypothesis agrees with the [Fernández-García et al. 2021] study in which the inclusion of social and demographic data in the models was suggested.

After selecting Model 2, we conducted three experimental repetitions for each classifier to determine if there were any statistical differences among the algorithms. Based on the statistical analysis, the RF, GB, and SVM algorithms performed the best. We chose to present the results of the GB algorithm and plan to evaluate another GB-based algorithm, XGBoost, in future work.

We analyzed Model 2 with the GB algorithm, which achieved F1-Score scores between 69% and

85%, depending on the dataset. For most models and classifiers, Dataset A consistently yielded the weakest results across most models and classifiers, while Dataset D consistently yielded the highest results. This outcome was expected due to the varying number of features available in each dataset. Thus, we concluded that predicting dropout in the third semester (Dataset A) is not as reliable as predicting in the ninth semester (Dataset D). The prediction probability of predictions performed correctly reached ranges between 90% and 100% in Dataset D.

In conclusion, this study successfully identified students who are more likely to drop out and identified the main features that contribute to dropout prediction. The results of this study consistently showed that the final grades of the disciplines were the most influential feature in predicting student dropout. By employing machine learning techniques and analyzing academic data, we were able to predict dropout tendencies with reasonable accuracy. The findings of this study can assist educational institutions in implementing targeted interventions and support systems to prevent student dropout and improve overall retention rates.

REFERENCES

Alban, M. and Mauricio, D. Predicting university dropout through data mining: A systematic literature. *Indian Journal of Science and Technology* 12 (4): 1–12, 2019.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

Breiman, L. Random forests. *Machine Learning* vol. 45, pp. 5–32, 2001.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and Regression Trees*. CRC Press, 1984.

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. Choosing multiple parameters for support vector machines. *Machine Learning* 46 (1-3): 131 – 159, 2002.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, pp. 785–794, 2016.

Fernández-García, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., and Sánchez-Figueroa, F. A real-life machine learning experience for predicting university dropout at different stages using academic data. *IEEE Access* vol. 9, pp. 133076–133090, 2021.

Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29 (5): 1189–1232, 2001.

Geurts, P., Ernst, D., and Wehenkel, L. Extremely randomized trees. *Machine Learning* vol. 63, pp. 3–42, 2006.

He, H. and Ma, Y. *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, 2013.

INEP. Brazilian higher education census. https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior, 2022.

Ishwaran, H. The effect of splitting on random forests. *Machine Learning* 99 (1): 75–118, 2015.

Romero, C. and Ventura, S. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery* 10 (3): e1355, 2020.

Santos, C. H. D. C., de L. Martins, S., and Plastino, A. Is it possible to predict dropout based on academic performance only? *Brazilian Symposium on Informatics in Education* vol. 32, pp. 792–802, 2021.

Santos, G. A. S., Bordignon, A. L., Oliveira, S. L. G., Haddad, D. B., Brandão, D. N., and Belloze, K. T. A brief review about educational data mining applied to predict student's dropout. In *Anais da V Escola Regional de Sistemas de Informação do Rio de Janeiro*. SBC, Porto Alegre, RS, Brasil, pp. 86–91, 2018.

UFPR. Bachelor's degree in computer science - curricular grade. https://web.inf.ufpr.br/bcc/curriculo/grade-curricular-2011/, 2011.

UFPR. Previous entries. https://servicos.nc.ufpr.br, 2022.

Zhu, J., Zou, H., Rosset, S., and Hastie, T. Multi-class adaboost. *Statistics and Its Interface* 2 (3): 349–360, 2009.