Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Dayane Perez Bravo

Evaluating Strategies to Predict the Evasion of Students in Higher Education

Curitiba
2022

Dayane Perez Bravo

Evaluating Strategies to Predict the Evasion of Students in Higher Education

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Marco A. Zanata Alves

Curitiba
2022

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Evaluating Strategies to Predict the Evasion of Students in Higher Education

Dayane Perez Bravo[1]
Leandro Augusto Ensina[2]
Luiz Eduardo Soares de Oliveira[3]
Marco Antonio Zanata Alves[3]

Abstract

The Brazilian Higher Education Census showed that the dropout rates of higher education students in Brazil exceed 50% from the fifth year onwards. This high tax of evasion causes several problems in terms of wasted resources invested by the society and the student. Therefore, any university must develop strategies to prevent student dropout and minimize those problems. Nevertheless, predicting student evasion involves detecting patterns and predicting them over a high amount of data collected yearly from thousands of students. Considering the dimension and the amount of data involved in dropout prediction, one can suggest using Machine Learning techniques to automate the identification of these students. This paper aims to identify dropout-prone students based on the behavior history of students at an unpaid public university. We engineered four datasets according to the semester in which the student is in the course. Such datasets intend to simulate the academic scenario and individual features of the students available until the moment of the prediction. Statistical tests showed a significant difference between the three feature models proposed. Our method could identify the students most likely to drop out and their main characteristics. 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.
Keywords: Pattern Recognition, Feature Engineering, Student Dropout.

Resumo

O Censo do Ensino Superior brasileiro mostrou que as taxas de evasão dos estudantes do ensino superior no Brasil ultrapassam 50% a partir do quinto ano. Essa alta taxa de evasão causa diversos problemas em termos de desperdício de recursos investidos pela sociedade e pelo aluno. Portanto, qualquer universidade deve desenvolver estratégias para evitar a evasão de alunos e minimizar esses problemas. No entanto, prever a evasão estudantil envolve detectar padrões e prevê-los em uma grande quantidade de dados coletados anualmente de milhares de alunos. Considerando a dimensão e a quantidade de dados envolvidos na previsão de evasão, pode-se sugerir o uso de técnicas de Machine Learning para automatizar a identificação desses alunos. Este trabalho tem como objetivo identificar alunos propensos à evasão com base no histórico do comportamento de alunos de uma universidade pública gratuita. Projetamos quatro conjuntos de dados de acordo com o semestre em que o aluno está no curso. Tais conjuntos de dados pretendem simular o cenário acadêmico e as características individuais dos alunos disponíveis até o momento da previsão. Testes estatísticos mostraram uma diferença significativa entre os três modelos de características propostos. Nosso método conseguiu identificar os alunos com maior probabilidade de evasão e suas principais características. Utilizar apenas as informações das disciplinas cursadas pelos alunos mostrou-se o melhor modelo de características. Ao usar esse modelo com o Gradient-Boosting, o desempenho do F1-Score variou entre 69% e 85%, dependendo do conjunto de dados.
***Palavras-chave**: Reconhecimento de Padrões, Engenharia de Atributos, Evasão de Estudantes.*

1 Introduction

According to the Brazilian Institute of Research on Education (Inep) [1], dropout is characterized by the student leaving a course before completing it, regardless of the reason. Several situations can be characterized as a dropout. Among these situations, the lack of attendance, abandonment, dismissal, and transfer to another course or educational institution are examples. When a student drops out, the main consequences are financial and social [2]. Financially, one can cite the commitment of the budget of educational institutions that will not have good use of the physical space, equipment, and staff, since the classrooms are empty during the course. From the social point of view, multiple job opportunities lack qualified people, and individuals have fewer chances of improving prosperity.

The Brazilian Higher Education Census carried out

[1]Student from the Data Science & Big Data Specialization Course, dayaneperezbravo@hotmail.com.
[2]Student from Department of Informatics - UFPR, leandro.ensina@ufpr.br.
[3]Professors from Department of Informatics - UFPR, lesoliveira@inf.ufpr.com and mazalves@inf.ufpr.com.

by Inep [3] between 2011 and 2020 observed the permanence of those entering undergraduate courses in Brazil. Keeping the year and entry course fixed, the census used three concepts: (i) **Permanence:** when the student remains active/enrolled in his entry course; (ii) **Dropout:** when the student leaves the entrance course; (iii) **Completion:** when the student graduates from his entrance course. At the end of the first five years of follow-up, 51% of the entrants dropped out, and 29% completed the course, as shown in Fig. 1. At the end of the ten years of follow-up, 40% of the entrants completed the course, while 59% dropped out. In this way, we can glimpse the importance of creating policies for student retention.

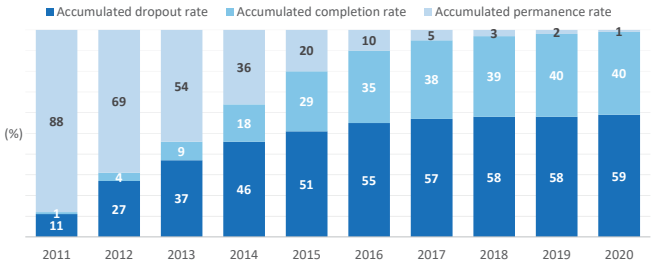


Figure 1: Evolution of the trajectory of those entering undergraduate courses in Brazil between 2011 and 2020. Adapted from [3].

According to a consultation carried out on the Brazilian Ministry of Education website (e-MEC) [4], in Brazil, the Bachelor's Degree in Computer Science (BCC) is offered by 431 Higher Education Institutions (HEI), in which 65% of these institutions are private, and 35% are public. The offer of vacancies by the BCC among private and public institutions represents, respectively, 82% and 18% of the annual vacancies offered in the country. Such undergraduate courses in Brazil have an average duration of five years. More than half of the students dropped out of their courses within this period, according to the Inep census [3].

With the development of student retention policies in higher education, the damage caused by dropouts could be reduced. Therefore, it is relevant to understand the causes of dropout and identify students most likely to drop out of their courses. These predictions shall allow HEI to develop actions that encourage them to remain in the course. However, such prediction may involve a high amount of data considering the dimension of the problem. Thus the use of Machine Learning (ML) techniques is recommended to automate the identification of these individuals [5, 6].

In this work, we aim to identify dropout-prone students using ML techniques and the essential characteristics to achieve such predictions. We propose to use a database with the behavior history of students of a BCC along the course from an unpaid public university in Brazil, the Universidade Federal do Paraná (UFPR). We engineered the original database into four datasets according to the semester of the course (3rd, 5th, 7th, and 9th) in which the student is. These datasets intend to simulate different scenarios for the students during their courses, virtually isolating only the available information until the moment of the prediction. Moreover, we developed three characteristic models to discover the principal characteristics that better predict dropout-prone students.

The main contributions of this work are: (i) feature engineering performed for the construction of datasets capable of representing a real scenario; (ii) performance evaluation through the AUC indicated for unbalanced data; (iii) identification of the most important features for prediction; (iv) identification of the dropout-prone students; and (v) obtaining a specific model for a BCC course from a public and tuition-free 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 & Ventura [7] stated that Educational Data Mining (EDM) aims to assist efforts in educational institutions based on available data, ML techniques, and understanding of the educational management system. They noted that previous work applied the EDM between 2000 and 2018 to predict student dropout. In comparison, we propose using ML to predict dropout-prone students.

Alban & Maucirio [5] published a survey showing that classifiers based on decision trees represented 79% of the research on student dropout prediction published between 2006 and 2018. Fernández-García *et al.* [8] used demographic and academic data to predict students most likely to drop out. They tested Support Vector Machine (SVM), Gradient Boosting, and Random Forest classifiers in five models. They engineered the first model as the moment of enrollment, and the others as the first four semesters in which the student is. They reported an accuracy ranging between 70 and 91%, achieving the lowest score with the first model and the highest with the fifth. In contrast, we rely on different semester periods in our models, using decision trees, among other classification algorithms.

According to Rai & Jain [9], the main features that cause dropout are personal (28%) and learning difficulties (10%), through the results obtained by the ID3 and J48 classifiers. For Brito *et al.* [10], the performance in the first semester disciplines influenced dropout with an accuracy of over 70%. For Santos *et al.* [11], models based on decision trees obtained accuracy between 79.31% and 98.25% and could predict the dropout of students from any other educational institution. The authors used for prediction only the academic performance data of the students in each semester of the course.

Although these three previous works aimed to predict the dropout rate of students in Computer Science courses in the same way as the present research, in this work, we propose three feature models: the first with only the students' features; the second with only the students' academic features; and the third with both features.

3 Proposed Method

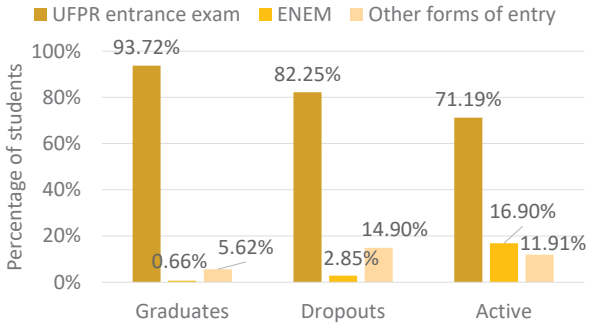
The database used for this study contains the anonymous information of the 2,763 students who entered the BCC course at UFPR between 1995 and 2019. The data on students in this database are the following: gender, year of the curriculum, forms of admission and evasion, year and period in which admission and evasion occurred. In addition, the information on disciplines taken by each of these students are: code, name, final grade, year and period in which they attended, and approval status. Based on this original database, we extracted some information as follows.

3.1 Data Understanding

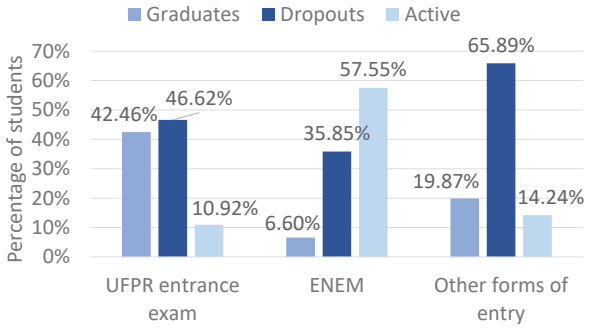
The behavior of the ways of dropout concerning the entryways analyzed in this study is illustrated in Fig. 2(a). Among the students who graduated, 93.72% entered through the UFPR entrance exam, 0.66% through Brazilian National High School Exam (ENEM), and 5.62% through other entryways. The relation between the students' enrollment and dropout analyzed in this study is illustrated in Fig. 2(b). Among the students who entered through the UFPR entrance exam, 46.62% dropped out, and 42.46% graduated. In addition, 10.92% of these students remain active in the course.

The behavior of admission forms concerning student gender is present in Fig. 2(c). Among the students who entered through the UFPR entrance exam, 11.59% were female and 88.41% male. Among the three forms of admission analyzed, all of them have the majority of students concentrated in the male gender. The behavior of gender concerning the ways of dropout is shown in Fig. 2(d). Among female students, 45.75% dropped out, and 43% graduated, where the percentage of female graduates is higher than the male gender.

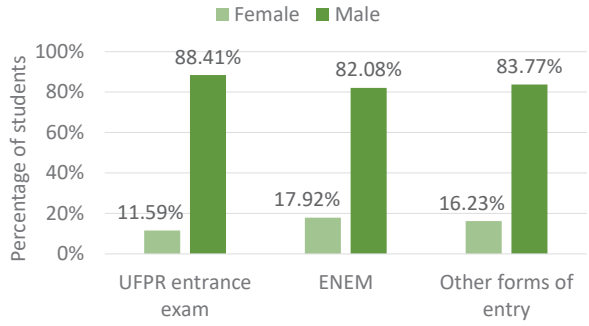
We calculated the students' dropout semester during the data cleaning process (Section 3.2), as presented in Fig. 3. It illustrates the behavior of the ways of dropouts concerning the semester in which the student is enrolled for the selected database. The number of dropouts in each semester diverges between 10 and 20 students. For example, the BCC course at UFPR lasts eight semesters, but only three students completed the course until this period for the selected database. The peak of graduated students occurs in the 11th semester. Based on this preliminary information, we started selecting and organizing the data.



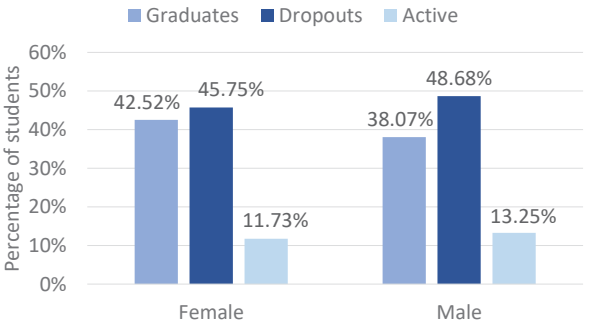
(a) Ways of dropout by entryways.



(b) Entryways by ways of dropout.



(c) Entryways by student gender.



(d) Ways of dropout by student gender.

Figure 2: Exploratory data analysis.

3.2 Data Preparation

Throughout the history of the BCC, several curricula were adopted according to the university's and students' best interests. The most recent curriculum changes were made in 2011 and 2019. Students over the 2011 curriculum were selected from the database to ensure that all students had the same mandatory disciplines. These dis-

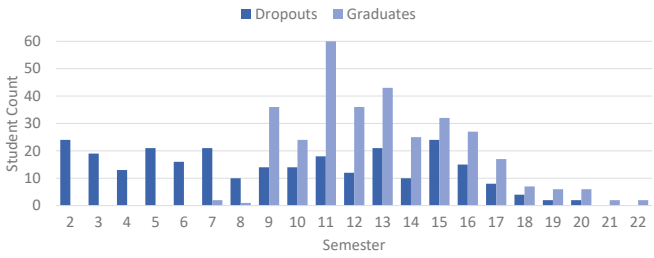


Figure 3: Number of graduates and dropouts by semester.

ciplines and the respective periods in which they were offered are available in reference [12]. Students who remain active in the course were removed because it is unknown if they will drop out or graduate. Notice that data from students' situations were collected in 2019, thus presenting a picture of that moment.

The years between 2015 and 2018 were not selected because the percentage of students still active in the course was higher than 30%. Thus, the years of entry between 2006 and 2014 were selected for the study, as illustrated in Fig. 4. In this context, we notice that students commonly migrate to newer curricula as soon as they are allowed. This migration explains why students entering before 2011 are enrolled with a newer curriculum (also justifying our selection of students with an entry year lower than 2011). Furthermore, disciplines that do not belong to the regular course curriculum were removed, leaving 594 students in the final database, in which 326 graduated and 268 dropped out.

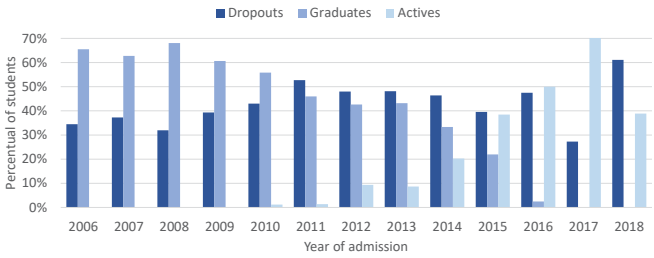


Figure 4: The proportion of students active, graduate and dropout in 2019 divided by admission year.

As discussed in the previous paragraph, we performed a database cleaning with selected data, removing, for example, special characters and irrelevant columns or columns with repeated information. These columns are: student registry code; course name; course code; curriculum code; name of disciplines; disciplines theoretical and practical hours. Based on the available information, some new attributes were calculated for each student: (i) the semester of the course in which he took each discipline; (ii) the semester in which the student dropped out; (iii) how many times the student took each discipline; and (iv) whether the student was periodized concerning the regular schedule. According to the year of entry of each student, the candidate/vacancy ratio was inserted, available by the UFPR agency responsible for the entrance selection exam [13].

Furthermore, five columns were inserted to simplify the information regarding: (i) gender; (ii) belonging of the discipline to the regular schedule; (iii) the final status of the discipline; (iv) evasion; and (v) admission forms. Moreover, the predominant classes were filled with '1' and the others with '0' as described below. Filled with '1' are: (i) male gender; (ii) disciplines that belong to the regular schedule; (iii) approved in the discipline; (iv) evasion through graduation; and (v) UFPR entrance exam. Filled with '0' are: (i) female gender; (ii) disciplines that do not belong to the regular schedule; (iii) failed in the discipline; (iv) evasion through withdrawal, cancellation, abandonment, transfer, and so on; and (v) other entryways like ENEM, transfer from another HEI, among others.

So, we have two sets with distinct information: demographic and academic. The former contains the following information: gender, evasion and admission forms, the semester in which the evasion occurred, and candidate/vacancy ratio. The latter includes the following information: discipline code, final grade, belonging of the discipline to the regular schedule, the semester of the course in which each discipline was taken, how many times the student took each discipline, and whether periodized concerning the regular schedule.

3.3 Data Manipulation

We created a table where all the features from the student appear in a single row. Each feature had a fixed position in the columns of this table. The data of the disciplines were placed side by side. When the student took the discipline more than once, it was selected only data from the last time the discipline was taken. We consider a given student periodized when all the disciplines are taken on the same semester described in the regular curriculum or whenever it was anticipated. A student is not periodized whenever some discipline is taken later or there is reprobation, e.g., when the student takes the discipline for the first time in the regular semester and the next time in a further semester.

We created four datasets (A, B, C, and D) simulating the scenario of the data available at the course's beginning of semesters 3, 5, 7, and 9. For **Dataset A**, were selected only the students whose dropout semester was higher than or equal to 3. This is because Dataset A aims to predict the students who will drop out from the third semester onwards. Therefore, students who had already dropped out were removed. Similarly, **Datasets B, C, and D** were created with only the students whose dropout semester was higher than or equal to 5, 7, and 9, respectively.

For Dataset A, were selected only data from mandatory disciplines whose semester was lower than 3. It was made because data from other disciplines could cause a bias in the model, as only a few students had anticipated disciplines. That is, it was removed data

from disciplines with a regular semester greater than or equal to 3. Datasets B, C, and D selected only data from mandatory disciplines with semesters less than 5, 7, and 9, respectively.

An example of a selection of the datasets is shown in Table 1. It simulates the data of three students. According to the regular schedule, we assumed that a D1 discipline should be taken in the second semester. Students X and Z took this discipline only once, while student Y took this discipline twice. For this same discipline, Student Y graded 45 the first time and 85 the last time (in the 4th semester).

Student Z dropped out in the 2nd semester; thus, we eliminated this student from Dataset A, leaving only students X and Y. Student X's D1 data was removed because the semester in which he took D1 was higher than 2. Likewise, the data from the second time that student Y took D1 was eliminated, as this data would only be available in semester 4. So Dataset A results are shown in the center of Table 1.

As an example of the selection of Dataset B, student Z was eliminated because his dropout semester was less than 5. However, the D1 data of student X was kept. The second time that student Y took D1 was a semester lower than 5, so this data was selected. Thus, the data from the first time was removed. Dataset B results are shown on the left of Table 1.

3.3.1 Data Splitting

According to each of the four generated datasets, an analysis was made between the rate of dropouts and graduates. This selection was made to ensure that rates were maintained. So, for the training and testing datasets, the rates of graduates and dropouts from the initial datasets were the same. For each dataset, 80% of the students were randomly selected for the training phase and the other 20% for the testing. We strictly used only the training data during the development of the models. From the training dataset, a new division was made. We randomly selected 20% of these students for the validation dataset and 80% for the training. The testing datasets were used at the end of this process to validate and report the results.

3.4 Models and Evaluation Methods

We designed three attribute models in this research. **Model 1** with only three personal characteristics of each student: gender, admission form, and candidate/vacancy ratio; **Model 2** with only four features referring to the disciplines taken by the students: final grade, the semester (within the course) in which each discipline was taken, how many times each discipline was taken and whether periodized concerning the regular schedule; and **Model 3** combines both features of the previous models, 3 for each student and 4 for each discipline.

As each dataset has a different number of disciplines, the number of attributes for each dataset is also different. For Model 2, Dataset A has 10 mandatory disciplines, as we have four features for each discipline, we get a total of 40 attributes. For Dataset B, we have 20 mandatory disciplines, therefore, 80 attributes. For Dataset C, we have 30 mandatory disciplines, therefore, 120 attributes. For Dataset D, we have 34 mandatory disciplines, therefore, 136 attributes. Now, for Model 3, the calculation for the number of features per discipline is the same as in Model 2. To obtain the total number of attributes of each dataset of Model 3, just incorporate the features of both models previously described. Thus, we have that Dataset A has 43 attributes, B has 83, C has 123, and Dataset D has 139 features.

We chose supervised learning algorithms as the data in this research have a label. These labels have two possible classes: graduate or dropout. Thus, there is a classification problem. The seven classifiers implemented were: (i) Decision Tree (DT) [14]; (ii) Extra-Trees (ET), with 100 estimators [15]; (iii) Random Forest (RF), with 100 estimators [15]; (iv) Gradient Boosting (GB), with 100 estimators [14]; (v) AdaBoost (AB), with 50 estimators [14]; (vi) Support Vector Machine (SVM) [14]; (vii) Logistic Regression (LR) [16].

To understand the feature's importance and predict probability, we used two methods in the scikit-learn library (version 1.0.2) [17]. The **feature importance** was used to identify which features were significant to the model's performance. The classifiers that have this measure are: ET, RF, GB, and AB. The **predict probability** was used to identify the probability of each student dropping out of the course. It measures the probability of the instance being classified among the problem classes, namely graduate or dropout.

The classes in our dataset are imbalanced, as the number of graduates and dropouts are different. Because of this, we selected the metric Area Under the Curve (AUC) and F1-Score [18].

4 Results and Discussion

For training, the attributes form of evasion and semester in which the evasion occurred were not considered since this information would not be available in a real scenario. The performances of the seven classifiers were evaluated for each of the three models according to the four datasets. It is worth mentioning that some original features in the datasets were removed during the validation step since they did not demonstrate improvement in the algorithms' performances (less than 2%). The features eliminated during these analyses were: if discipline belongs to the regular schedule, how many times each discipline was taken, and whether it periodized concerning the regular schedule. So the remaining features were: code, final grade, and the semester of the course in which each discipline was taken. After this feature selection, the models were re-trained and the AUC performances obtained are shown in Table 2.

Now, we present the statistical tests performed to verify if there is a statistical difference between the models

Table 1: Example of student data.

	Dataset Initial			**Dataset A**		**Dataset B**	
	Students			Students		Students	
	X	Y	Z	X	Y	X	Y
Semester in which the student dropped out	5	5	2	5	5	5	5
How many times D1 was taken	1	2	1	0	1	1	2
Last semester that D1 was taken	3	4	2	0	2	3	4
Last D1 Grade	90	85	60	0	45	90	85

Table 2: Classifiers performances: AUC.

	Data set	**Ada Boost**	**Random Forest**	**Gradient Boosting**	**Logistic Regression**	**SVM**	**Extra Trees**	**Decision Tree**
Model 1	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
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	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

and the classifiers. For all statistical tests, we used a significance level of 5%. Through Friedman's test for AUC, we found that the models are statistically different from each other. The analysis of rankings observed that: (a) Model 1 is significantly different from Models 2 and 3; and (b) Models 2 and 3 are statistically equivalents.

Model 1 has only three features, and, in addition, its performance averages (i.e., 58%) were lower than the other models (i.e., greater than 74%). Therefore, the statistical difference shows that Models 2 and 3 are better than Model 1, which can be explained because they have more features (information) than Model 01. On the other hand, Models 2 and 3 have no statistical difference, so we considered that Model 2 is better, as it requires fewer attributes than Model 3.

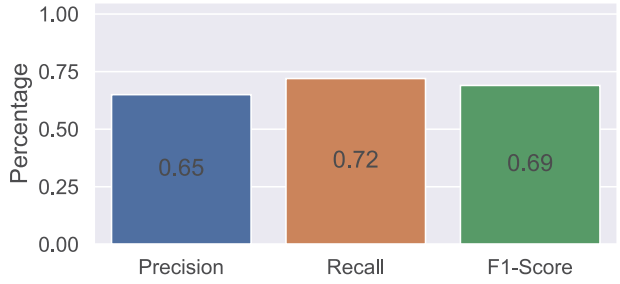
With the selection of Model 2, the training was repeated three times to test the statistical difference between the classifiers. Through Friedman's test, the RF, GB and SVM algorithms had the best performances (i.e., greater than 78%). Of these three, the RF and the GB allow identifying the attributes with the greatest influence on the predictions, an aspect that SVM does not allow. Consequently, we are left with just RF and GB.

We chose to present the GB results because we intend in future works to evaluate another algorithm that is based on GB, the XGBoost [19]. So the performance of GB applied to Model 2 is shown in (i) Fig. 5 for dataset A; (ii) Fig. 6 for dataset B; (iii) Fig. 7 for dataset C; and (iv) Fig. 8 for dataset D.

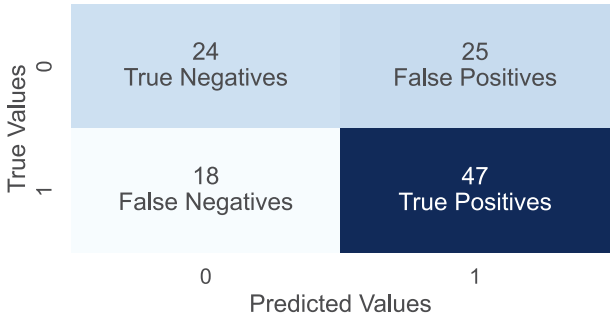
Analyzing the performance of GB in Dataset A concerning the confusion matrix, there are 71 successes and 43 errors (Fig. 5(b)). The AUC score was 67%, the lowest among the other datasets (Fig. 5(c)). Most successes occurred with probabilities between 68% and 90%, approximately (Fig. 5(d)). For this dataset, the three most important features were the final grades of the disciplines: (i) "*Introduction to Algebra*", with 24% of significance; (ii) "*Algorithms and Data Structures I*", with 17%; and (iii) "*Analytical Geometry*", with 15% of significance. These three disciplines are offered in the first semester.

The next analyzed performance of GB is in Dataset B, where there are 75 successes and 33 errors (Fig. 6(b)). The AUC score was 71% (Fig. 6(c)), and most successes occurred with probabilities between 90% and 93% (Fig. 6(d)). For this dataset, the three most important features were the final grades of the disciplines: (i) "*Digital Projects and Microprocessors*", with 30% of significance; (ii) "*Introduction to Algebra*", with 9%; and (iii) "*Algorithms and Data Structures I*", with 7% of significance. These three disciplines are offered in the first semester.

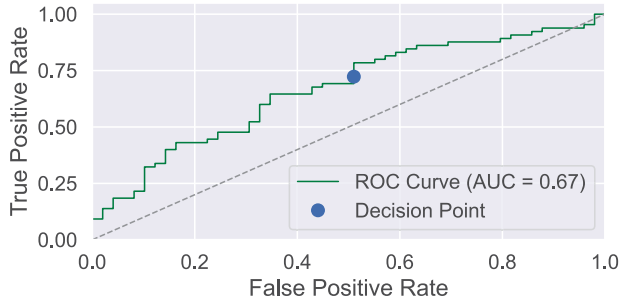
Now, the performance of GB in Dataset C shows that there are 75 successes and 25 errors (Fig. 7(b)). The AUC score was 85% (Fig. 7(c)), and most successes occurred with probabilities between 90% and 100% (Fig. 7(d)). For this dataset, the three most important features were the final grades of the disciplines: (i) "*Basic Software I*", with 19% of significance; (ii) "*Discrete Mathematics*", with 15%; and (iii) "*Computer Organization and Architecture*", with 7% of significance. These three disciplines are offered in



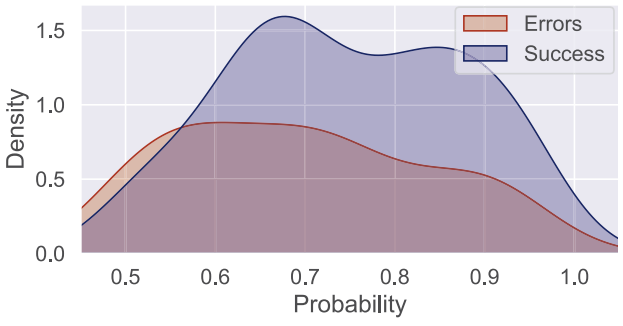
(a) Scores.



(b) Confusion matrix.

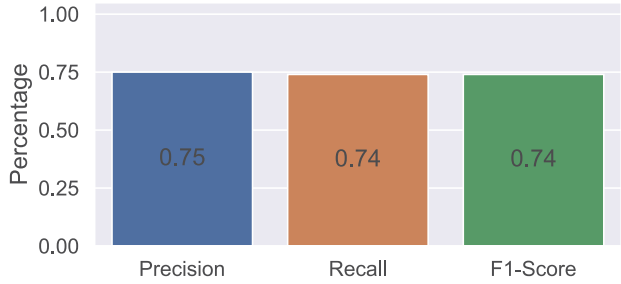


(c) ROC curve.

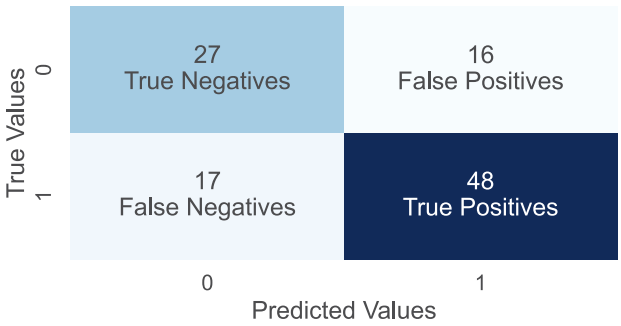


(d) Predict probability.

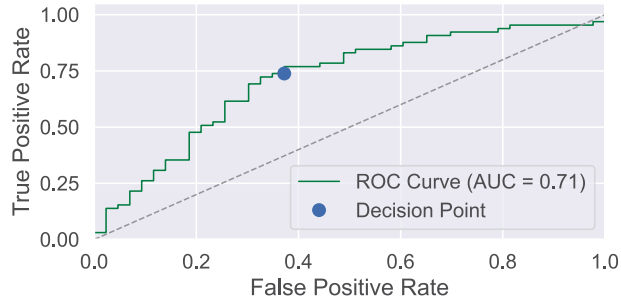
Figure 5: Gradient Boosting: Dataset A.



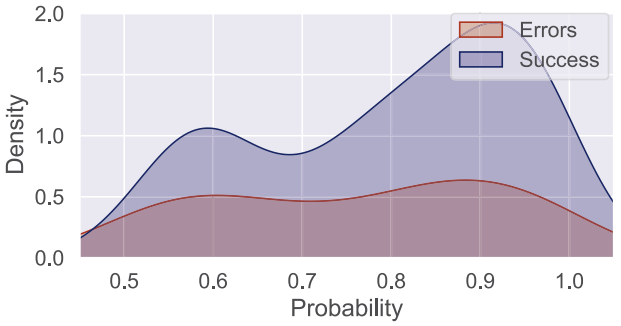
(a) Scores.



(b) Confusion matrix.



(c) ROC curve.



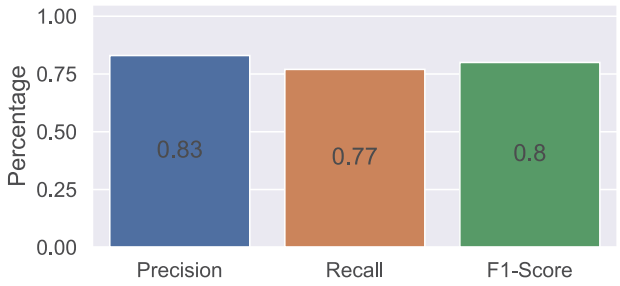
(d) Predict probability.

Figure 6: Gradient Boosting: Dataset B.

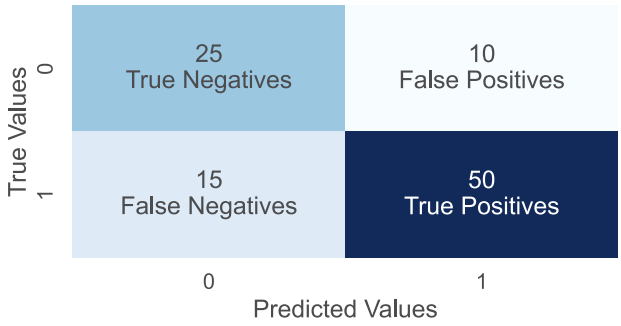
the third semester.

Finally, the performance of GB in Dataset D shows that there are 75 successes and 18 errors (Fig. 8(b)). The AUC score was 91%, the highest among the other datasets (Fig. 8(c)). Most successes occurred with probabilities between 90% and 100% (Fig. 8(d)). For this dataset, the three most important features were the final grades of the disciplines: (i) "*Discrete Mathematics*", with 36% of significance 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 significance and offered in the third semester.

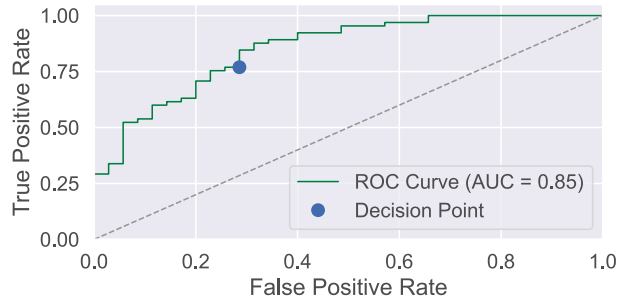
As we added information for more semesters, we could observe an improvement in the performance of the AUC and F1-Score. We expected this improvement in the performance as more semesters are used since more information about students is provided to the algorithms, and, consequently, greater discriminative power for classification. This behavior explains the lower per-



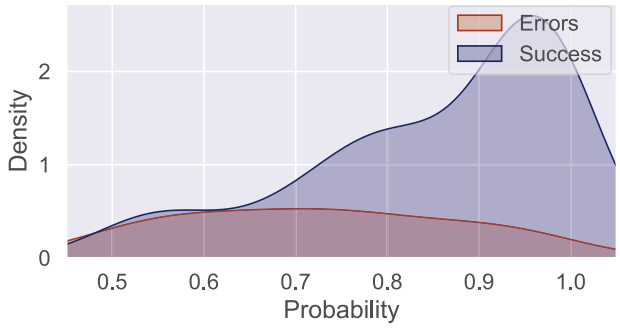
(a) Scores.



(b) Confusion matrix.

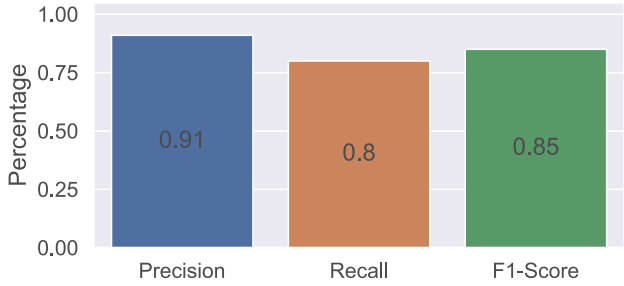


(c) ROC curve.

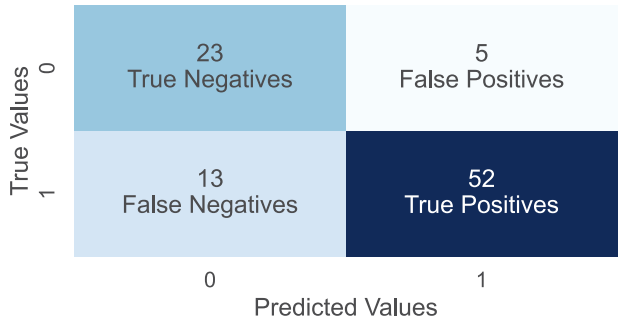


(d) Predict probability.

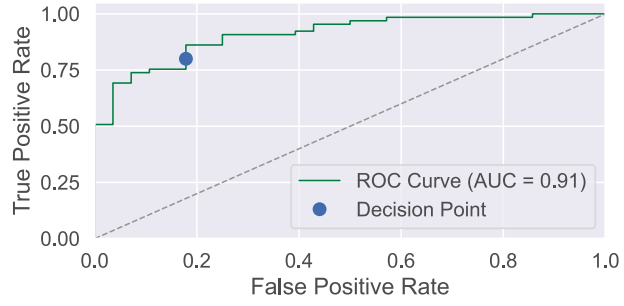
Figure 7: Gradient Boosting: Dataset C.



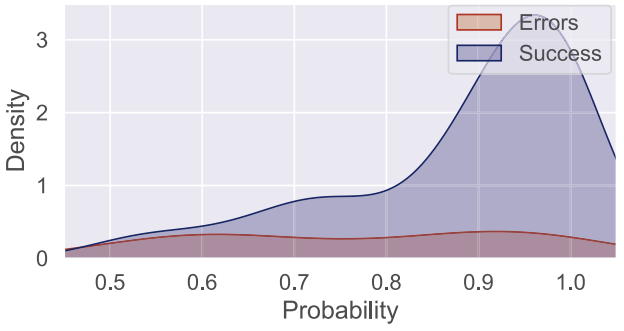
(a) Scores.



(b) Confusion matrix.



(c) ROC curve.



(d) Predict probability.

Figure 8: Gradient Boosting: Dataset D.

formance of Dataset A and the higher performance of Dataset D.

In addition, successes began accumulating at higher predict probabilities, while errors started to reduce in number at high probabilities. In this way, identifying students likely or not to drop out become substantial. 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%. This observation confirms the choice of Model 2 as the one with the best performance.

In a real-world scenario, datasets A, B, C, and D can be used simultaneously for different periods of the course. Education managers need to select 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 action on a case-by-case basis. These personalized action can be applied even with students who tend to stay in the course but with low 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. In addition, knowing the disciplines with the greatest influence, some actions can be taken, such as reinforcement classes and monitoring.

5 Conclusion

This paper proposes a method using ML to predict students prone to evading undergraduate courses. Working with different hypothetical scenarios formed by information from different university course phases, we could evaluate multiple algorithms and three attribute models. Separating these datasets was perceived as a way of simulating the real-world scenario.

Statistical tests were applied to the results obtained by the AUC metric. Through these tests, we concluded a statistical difference between Model 1 and the others (i.e., Models 2 and 3). Nevertheless, we considered Model 2 the best as it requires less data, and the features that contributed the most to the models were the final grade of the disciplines. Although the proposed method is directed to the BCC course, it can be adapted and evaluated for other courses (exact, biological, and human sciences), whether for a public or private institution.

Although Model 1 had the worst performance, we suggest in future work the inclusion of more personal data, such as: address, marital status, whether or not he/she is employed, family income, whether the student lives with their parents or not, number of children, among others. The low performance of Model 1 is probably because only three features were considered. A model with more personal data is likely to perform better. This hypothesis agrees with the research in which the inclusion of social and demographic data in the models was suggested [6].

After choosing Model 2, we conducted three experimental repetitions for each classifier to observe if there was a statistical difference between them. Statistically, the RF, GB, and SVM algorithms had the best performances. We chose to present the GB results because we intend in future works to evaluate another algorithm that is based on GB, the XGBoost. Evaluating other classifiers (e.g, XGBoost) in future works is a suggestion to improve these results.

Model 2 with the GB algorithm was analyzed, presenting the AUC scores between 67% and 91%, depending on the dataset. For most models and classifiers, Dataset A presented the weakest results, while Dataset D presented the highest. This result was expected due to the number of features available in each dataset. Therefore, we concluded that the prediction in the 3rd semester (Dataset A) is not as substantial as in the 9th semester (Dataset D).

By the end of this work, we could identify the students most likely to evade and the main features that induce this. The prediction probability of successes performed ranges between 90% and 100% in Dataset D. In all results, the final grade of the disciplines was the most crucial feature. In addition, we found that the disciplines of the first three semesters contribute most to the predictions. This shall indicate that educational managers should focus on the initial phases of the course. It is suggested in future work to include metrics to separate the disciplines into knowledge groups. This analysis can contribute to understanding any learning difficulty in certain areas of knowledge.

References

[1] INEP. Instituto Brasileiro de Estudos e Pesquisas Educacional Anísio Teixeira. Metodologia de cálculo dos indicadores de fluxo da educação superior, 2017. Last accessed 20 May 2022.

[2] G. Tontini and S. A. Walter. Pode-se identificar a propensão e reduzir a evasão de alunos?: ações estratégicas e resultados táticos para instituições de ensino superior. *Revista da Avaliação da Educação Superior*, 19(1):89–110, 2014.

[3] INEP. Instituto Brasileiro de Estudos e Pesquisas Educacional Anísio Teixeira. Censo da educação superior, c2022. Last accessed 11 Feb. 2022.

[4] eMEC Ministério da Educação. Cadastro nacional de cursos e instituições de educação superior, 2022. Last accessed 20 May 2022.

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

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

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

[8] Antonio Jesús Fernández-García, Juan Carlos Preciado, Fran Melchor, Roberto Rodriguez-Echeverria, José María Conejero, and Fernando Sánchez-Figueroa. A real-life machine learning experience for predicting university dropout at different stages using academic data. *IEEE Access*, 9:133076–133090, 2021.

[9] S. Rai and A. K. Jain. Students' dropout risk assessment in undergraduate courses of ict at residential university - a case study. *International Journal of Computer Applications*, 84(14):31–36, 2013.

[10] D. M. de Brito, I. A. de Almeida Júnior, E. V. Queiroga, and T. G. do Rêgo. Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. *Simpósio Brasileiro de Informática na Educação*, 25(1):882, 2014.

[11] C. H. D. C. Santos, S. de L. Martins, and A. Plastino. É possível prever evasão com base apenas no desempenho acadêmico? *Simpósio Brasileiro de Informática da Educação*, 32:792–802, 2021.

[12] Departamento de Informática – UFPR. Bacharelado em ciência da computação - grade curricular 2011, 2022. Last accessed 16 June 2022.

[13] Núcleo de Concursos UFPR. Vestibulares anteriores, 2022. Last accessed 20 May 2022.

[14] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, 2 edition, 2001.

[15] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:pages3–42, 2006.

[16] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[18] Haibo He and Yunqian Ma. *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, 1st edition, 2013.

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