Submission date: 21/08/2022
Resubmission date: 05/11/2022
Camera ready submission: 12/04/2023

Section: regular article

1st round notification: 06/10/2022
2nd round notification: 06/04/2023
Available online: 02/07/2023
Publication date: 02/07/2023

Improving the prediction of school dropout with the support of the semi-supervised learning approach

Eduardo Cardoso Melo[1], Fernanda Sumika Hojo de Souza[2]

[1] Departamento de Engenharia e Computação/Instituto Federal de Minas Gerais (IFMG)
Bambuí, Minas Gerais – Brasil

[2] Departamento de Computação/Universidade Federal de Ouro Preto (UFOP)
Ouro Preto, Minas Gerais – Brasil

eduardo.melo@ifmg.edu.br, fsumika@ufop.edu.br

Abstract. *School dropout is a phenomenon characterized by being influenced by several variables. This research used Machine Learning techniques, especially in the context of the semi-supervised learning strategy, to predict the risk of dropout in undergraduate courses at a Brazilian higher education institution. Two phases of experiments were conducted, the first using Feature Selection techniques and the second applying a semi-supervised learning strategy to improve performance metrics collected from the increase in the number of instances of students labeled as Graduated. As a main result, we obtained a model capable of classifying dropout with 90% accuracy and 86% Macro-F1.*

Keywords*. School Dropout; Machine Learning; Semi-supervised Learning; Educational Data Mining.*

1. Introduction

From the 1990s onwards, there was a significant expansion of private higher education institutions (HEIs) in Brazil. Of the 2,608 HEIs, only 302 (11.58%) are public (maintained by the State) (INEP, 2019). The government policy for Brazilian public higher education began to change from 2007 with the implementation of a support program for the restructuring and expanding of federal universities, named REUNI. This program allowed the emergence of new federal HEIs, the expansion of the number of campuses, courses and enrollments (Brasil, 2007; Neves & Martins, 2016), also favoring localities where there was no tuition free higher education. Importantly, the program also aimed at reducing the rate of dropout through the expansion of inclusion policies and student assistance.

With facilitated access to higher education, the profile of students in HEIs changed and began to involve individuals from different social and economic conditions, with new demands that, until then, were not part of the academic environment or were not priorities for administration (Nonato et al., 2020). In a scenario with high enrollment rates and a management that is still unable to meet the needs of students, school dropout has been impacted and requires specific studies both to understand its occurrence and to outline actions capable of mitigating it. As it is a phenomenon that suffers the effects of several variables, whether academic or personal, it becomes complex to build a broad model

Cite as: Melo, E. C. & Souza, F. S. H. (2023). Improving the prediction of school dropout with the support of the semi-supervised learning approach (Melhorando a previsão do abandono escolar com o apoio da abordagem de aprendizagem semi-supervisionada). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 16(1), 10:1-10:26. DOI: 10.5753/isys.2023.2852

capable of being adapted to the reality of each HEIs in Brazil. On the other hand, the history of school dropout mentioned in numerous studies published recently, together with the scenario brought about by the COVID-19 pandemic, in which many students had to drop out of studies or did not adapt to remote teaching, further increase the need for HEIs to broaden their horizons in relation to knowledge their audience. Therefore, detecting the factors that contribute to this phenomenon can improve the decision-making process by managers in the implementation of prevention policies (Neri & Osório, 2021).

School dropout causes several losses when it is carried out by the student, both in economic and social aspects. In a private institution, the drop in revenue can even compromise the course offering or, ultimately, the financial viability of the organization. In public institutions, the loss is for the whole of society, as their operation is funded by public resources earned through the collection of taxes. In addition, a considerable part of the budget matrix of each HEIs is conditioned to the achievement of indices related to the dropout and permanence of students in the institution, thus causing concern with the factors that can lead students to abandon their studies. Educational management itself tends to obtain benefits through specific studies on dropout in its environment, as the services and programs offered to the academic community can be created or adapted according to the demands of the institutional public, especially students. Having achieved knowledge about the main factors that contribute to dropout, it is possible that not only the management will start to act with prevention routines, but also the teachers themselves, who have a fundamental role in all processes related to the control of school dropout (Diniz & Goergen, 2019).

Recent studies have applied Machine Learning (ML) techniques to analyze dropout data in educational institutions, in order to understand the variables associated with this phenomenon as well as to build models capable of predicting new occurrences (Demeter et al., 2021; Musso, Hernández & Cascallar, 2020; Berka & Marek, 2021; Jia & Maloney, 2014; Perez, Castellanos & Correal, 2018; Agrusti, Bonavolontà & Mezzini, 2019). Previous researchs involving the construction of models for predicting school dropout with the support of ML techniques have used the supervised learning approach. Here, we have investigated whether the semi-supervised learning approach is able to improve the performance of school dropout prediction models in the context of undergraduate courses at a Brazilian federal HEI. Such a strategy allowed optimizing the performance of the models and also allowed the analysis and use of more than four thousand instances of students enrolled in order to outline the profile of those students who complete their courses.

This work contributes to the analysis of the phenomenon of school dropout from a different perspective, so that institutional managers can make use of the support of technologies and computer systems in their daily routines, such as the planning of actions that seek to eliminate dropout from the academic context or at least reduce the effects of student departure. Even if the main objective is not the elaboration and proposal of new techniques related to the ML area, its main contribution lies in the advancement of studies that use such resources, especially those linked to the semi-supervised learning approach, to improve predictions about school dropout and, consequently, increase the level of confidence in the use of the information generated as results.

2. Background

2.1. School dropout in higher education

School dropout involves all levels of education, it is not just a reality in higher education (Lee & Chung, 2019; Rumberger, 2020). However, more than identifying its causes and consequences, it is important to theoretically understand what this broad phenomenon is about, so that discussions are not based only on quantitative aspects, relegating other issues to the background (Nicoletti, 2019; Castro & Teixeira, 2013). In the understanding of Baggi & Lopes (2010), dropout refers to a stop in the study cycle, regardless of the student's level of education. Fialho & Prestes (2014) indicate that the understanding of dropout can be improved if this interruption of studies is interconnected with answers to how, when and why the student decided not to continue their course. According to these authors, HEIs need to minimally understand these points to be able to build strategies that allow them to reduce dropout rates in their environment. Some authors (Momm & Momm, 2020; José, Broilo & Andreoli, 2011) claim that dropout is already characterized even if the interruption is temporary, as it occurs, for example, when the course is suspended or when it ceases participation in classes and other school activities.

The causes of dropout can come from both the institution's internal and external environment, in addition to being related to the student's personal issues (Kehm, Larsen & Sommersel, 2019; Hedge & Prageeth, 2018; Pascoe, Hetrick & Parker, 2020; Biazus, 2004). Among the internal ones, there are situations involving didactic-pedagogical aspects (course curriculum, diversified academic activities, evaluation system), teaching staff (training, qualification and contact with students), student assistance (policy and socioeconomic support programs for permanence of students) and infrastructure of the institution (library, laboratories and adequate classrooms). In the external ones, issues related to the country's economic, political and social conditions are mentioned, which can compromise the student's viability to continue studying. For example, the worsening of the economic scenario can cause the student or his family group to lose the source of income that allows the student bond, causing the school dropout. As for the student's personal aspects, the authors indicate that family-related causes (health problems, early pregnancy or moving to another city) and lack of vocation (inadequacy of the chosen course or lack of bond with the chosen profession) can contribute to the interruption of studies (Balkis, 2018; Stadler et al., 2015; Silva Filho et al., 2007; Martins, 2007).

Dealing with the consequences, dropping out of studies generates several implications for personal issues of the student involved, such as feelings of frustration, failure, insecurity and intellectual incapacity, in addition to economic problems due to expenses already incurred, hired or even the lack of options in the job market for not having the required academic training (John et al., 2018; Koc, Zorbaz & Demirtas-Zorbaz, 2020; Nagai & Cardoso, 2017). In addition to the financial losses for the student or their family, the disruption of studies can lead to possible psychological damage if the reasons that led to dropout are not well addressed by those involved (Jagodics & Szabó, 2022; Castro & Teixeira, 2013). It is important to mention the aspect of social exclusion generated by the early termination of studies, as the study environment often provides possibilities never before seen by students with socioeconomic difficulties (Ceratti, 2008).

Regardless of the classification given to the interruption of bond by the student, it is essential that the situation is identified and analyzed so that it does not happen to other students for the same or similar reasons (Davok & Bernard, 2016; Shirasu & Arraes, 2016).

When the profile of public higher education students in Brazil is analyzed, the number of female students in HEIs has been systematically increasing, as well as individuals who declared themselves to be brown or black. As for family per capita income, the majority (66%) are in the eligibility range to claim socioeconomic assistance, which is up to one and a half minimum wage. The number of freshmen who attended high school in public schools (64%) also makes it possible to conclude that government policies to expand and diversify the forms of access to public higher education have had the desired effect (Andifes, 2019). Other factors contribute to the change in the profile of students, such as the access of the neediest individuals to educational institutions, the offer of new careers with a focus on practical work, the expansion of networked mobile technologies and the promotion of educational institutions to its educational services in traditional areas that do not require large investments in infrastructure (Diniz & Goergen, 2019).

2.2. Machine Learning

Machine Learning is an area that aims to build computational techniques on learning and to develop systems that have the ability to automatically obtain new knowledge, also enabling the development of new skills and supporting the organization of existing knowledge in various ways (Mitchell, 1997; Monard & Baranauskas, 2003). The ability of machines to learn in a context with the least possible amount of human intervention is fundamental for the existence of their intelligent behavior (Helm et al., 2020; Sousa, 2020; Batista, 2003).

Some features of Machine Learning can be better understood from a division of the concept into types. Ayodele (2010) proposes that Machine Learning algorithms be categorized into at least four types: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. In supervised learning, labeled data is used in the training stage, i.e., data is associated with a target class (Monard & Baranauskas, 2003; Morais et al., 2020). Input data is iteratively adjusted as it is inserted until it becomes properly adjusted with minimal error, allowing it to be able to produce accurate results when presented with new data. This allows the algorithm to understand and, mainly, learn how that result was obtained, so that it can then try to predict the output in another dataset whose target class value has not yet been defined for each instance (Jordan & Mitchell, 2015). Unsupervised learning is applied to unlabeled data, i.e., no category information is previously available in the dataset for use during training activities. It involves the application of techniques capable of analyzing the entire dataset and, based on various observations, indicating a possible category for each instance. Algorithms check for patterns in the dataset and are able to solve clustering or association problems (Fisher, Pazzani & Langley, 2014). Reinforcement learning seeks to enhance learning by gaining new experiences through trial and error. The results are obtained from the training of an agent that interacts with the environment in order to cumulatively increase the rewards and, thus, reinforce the learning built so far (Wang & Taylor, 2017).

Semi-supervised learning is characterized as an intermediate type between supervised and unsupervised learning (Gibson, Rogers & Zhu, 2013), so that the dataset will have both labeled and unlabeled data at the same time. It is unlikely that a fully supervised or totally unsupervised approach will be able to adapt to heterogeneous scenarios, as we find in the real world (Van Engelen & Hoos, 2020), and therefore, semi-supervised learning may have an interesting applicability. One of the possibilities of applying the algorithms of this type of learning is to train the classifiers using the labeled data to, in a second moment, predict the target class of the unlabeled data, thus generating the so-called pseudo-labeled instances. Then, the pseudo-labeled data obtained are combined with the labeled data for retraining the classification model, such as supervised learning (Guo, Wang & Li, 2021). According to Zhu & Goldberg (2009), the motivation behind semi-supervised learning classification lies in training a classifier from both the labeled and unlabeled data, resulting in a better performance than the supervised learning classifier on the labeled data alone. One popular approach is called self-training models, whose main characteristic lies in the fact that the learning process uses its own predictions to teach itself, having the advantage of the simplicity of building the model combined with the possibility of integration with various classification algorithms, from simple to complex ones.

2.3. Related works

Teodoro and Kappel (2020) studied the phenomenon of school dropout in the context of Brazilian HEIs, aiming to identify the most determinant characteristics for students to drop out and, thus, try to predict the possible dropout of other students with similar characteristics. Five ML algorithms (Naive Bayes, K-Nearest Neighbors, Decision Trees, Random Forest and Neural Networks) were applied to a dataset obtained from INEP. As a main result, the study indicated that the dropout of HEIs students is more related to age, with the total workload of the chosen course and with eventual participation in extracurricular activities. The large number of instances (376,746) used in the process of creating predictor models stands out in this work. The Random Forest and Neural Networks algorithms presented the best performance in terms of accuracy (80%) and Macro-F1 (79%) as predictor models.

The dropout rate of undergraduate students was analyzed by Gonçalvez, Silva and Cortes (2018) to identify those with a tendency to leave the institution. Based on the methodology proposed by the KDD process, the authors used the Naive Bayes, Support Vector Machine and J48 algorithms to build the learning models. As main results, the study pointed out that the J48 algorithm showed better accuracy, followed by SVM and, finally, by Naive Bayes. It is noteworthy, however, that the final average of the three algorithms was above 94%. Macro-F1 metric results data were not available. It is interesting to note that there were no significant differences between the classifications performed by the three algorithms, not even when Feature Selection techniques (Information Gain and Correlation Based Feature Selection) were applied. This situation can be explained by the extensive data pre-processing work carried out by the authors before the application of ML techniques.

Melo (2016) analyzed more than 32,000 instances of students enrolled in 76 higher education courses offered by an HEI and had as a differential the definition and comparison of two models for classifying students, one with all those linked to the institution and another segmented by course. Three Feature Selection techniques

(Information Gain, Gain Ratio and Symmetrical Uncertainty) were applied to both models. The work identified that it is possible to obtain satisfactory indications of which students tend to drop out using only academic data from the institution, regardless of the socioeconomic aspects of the students. A small set of attributes managed to obtain a performance similar to the set composed of all available attributes, indicating that the model can be optimized and promote greater interpretation capacity. Furthermore, the author presented evidence that the model with all students presented better results than the one segmented by course, correctly predicting the instances with 90% of accuracy and 80% of Macro-F1. All models were created using only the Random Forest algorithm, as observed in the study by Soares et al. (2020), whose results were improved after running the Grid Search process.

The study by Kantorski et al. (2016) has some similarities with the work proposed here, in particular the use of the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology as a support for the development of activities. Although small in terms of the number of instances (791), the same dataset was applied in two types of simulation, one considering the students labeled as graduated as regular students and the other without this consideration. The CART, J48 and Naive Bayes algorithms were used, with the best results coming from CART (73% accuracy and 46% Macro-F1). The simulation that integrated graduate students with regular students provided the best results in terms of success in predicting dropout, as it balanced the dataset in relation to the number of instances labeled as dropout.

The identification of the attributes most linked to the possibility of dropping out was one of the main objectives of the research by Assis (2017), which used data from an HEI together with data from the Census of Higher Education and the National High School Exam (ENEM). Although five different classification models were created based on Decision Trees, Artificial Neural Networks and Bayesian Methods approaches, the research did not find statistically significant variations between the metrics collected from the application of different algorithms. The best result was obtained by CART algorithm (92% accuracy and 53% Macro-F1). The activities developed were based on the CRISP-DM methodology. It was observed that the students most likely to drop out had as common characteristics the fact that they entered the first semester of the year, ENEM scores above average and a longer time between the completion of high school and the completion of the ENEM. It is interesting to note that the study presented comparisons between models created with and without balancing, not finding significant differences.

Berka and Marek (2021) used both data collected from students at the time of admission to the institution and academic data related to the first academic semester. The CRISP-DM methodology supported the organization of the study stages and the classification algorithms used were based on decision trees. Among the results presented, the low impact of the characteristics of the students at the time of their entry into the institution in the final classification as a graduate or dropout stands out. In addition, the analyzes showed that the longer the time between finishing high school and entering higher education, the greater the chances of dropping out.

Different classification models (Decision Tree, Bayesian Methods and Logistic Regression) were applied by Perez, Castellanos and Correal (2018) in an attempt to understand the main predictors of school dropout in the analyzed context, as well as to demonstrate that even algorithms considered as of simpler use offer satisfactory results

in terms of classification and prediction. The data set consisted of personal and academic information from 802 students who joined the institution between 2004 and 2010, with no data prior to their admission. The performance of the three models analyzed was relatively close in terms of accuracy: Decision Tree (94%), Logistic Regression (92%) and Bayesian Methods (87%).

3. Methods

The organization of the activities in this study were supported by the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, a reusable architecture in data analysis projects in companies from the most varied segments that has a standard process model based on iterative cycles (Wirth & Hipp, 2000). Figure 1 presents the standard structure proposed by CRISP-DM.

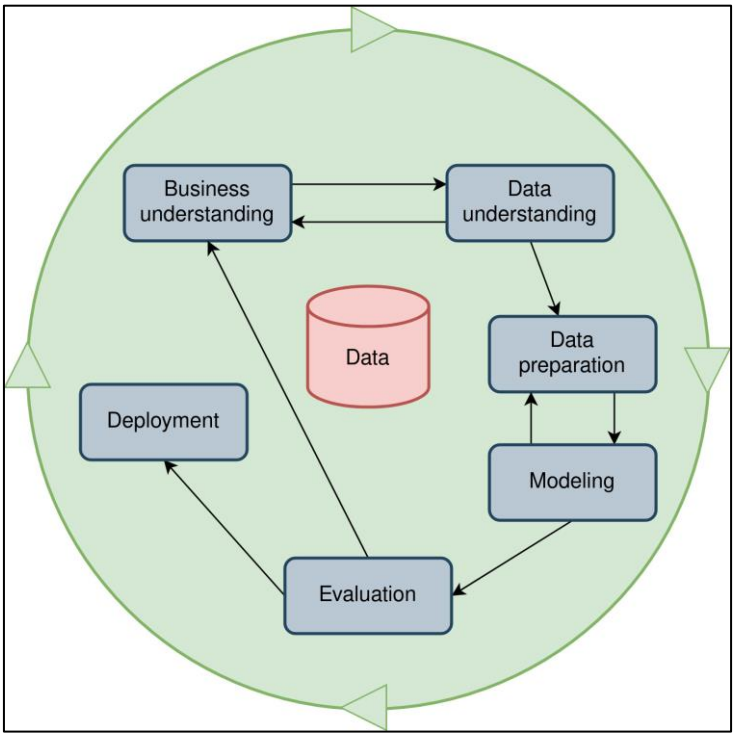

Figure 1. Stages of the CRISP-DM reference model

Source: Adapted from Niaksu (2015)

The first stage (*Business Understanding*) consisted of understanding the issues related to school dropout, especially how it has been conducted in the analyzed institution. It is a Federal Institution of Higher Education located in the state of Minas Gerais, Brazil, consisting of 18 campuses and with about 10,000 students currently enrolled, in which on-site higher education courses are offered in various areas of knowledge. The last study on school dropout at the institution was carried out in 2016 and showed that the percentage of students who dropped out of the course was high (~53%). It is important to mention that in this study are considered dropouts: students who dropped out of a course and remained at the institution in another course (internal transfer); students who dropped out of the institution (external transfer); students who have closed their enrollment;

students who were dismissed and considered dropouts by the institution; and students who dropped out of higher education.

In the second stage (*Data Understanding*) the structured data available within the institutional environment were analyzed, i. e. those that are currently maintained and treated with the support of information systems. Since 2013, the institution has been using an academic management system that stores both demographic and academic data of students, as well as the courses and subjects offered by the campuses. The anonymized data used in this study were made available by the institution in three files, one containing student records, another with records of the subjects studied by them and the last with data from the Student Assistance sector. We chose to analyze the data from the groups of students entering from 2013 to 2019, as we understand that the school dropout that occurred during the COVID-19 pandemic deserves a separate study, totally disconnected from the history that had been had until then.

The third stage (*Data Preparation*) involved the execution of activities to clean the data (duplicate student records, or without indication of current status), integration of demographic data with academic and socioeconomic data, as well as the transformation, aggregation and creation of new attributes. At the end of the activities of this stage, the dataset had 21 attributes (9 original and 12 created), which are named and detailed in Table 1.

Table 1. Dataset attributes

	Attribute	**Type**	**Created**	**Description/possible values**
01	year_entry	Numeric	Yes	Year of student entry into the course (2013 to 2019)
02	course_area	Categoric	Yes	Course Area (Agricultural Sciences, Biological Sciences, Exact and Earth Sciences, Human Sciences, Applied Social Sciences, Engineering, Linguistics, Letters and Arts)
03	campus	Categoric	No	Student campus (18 different cities)
04	course_worload	Numeric	No	Total course workload (1.600 to 4.887)
05	race_color	Categoric	No	White, Brown, Black, Indigenous
06	pc	Numeric	No	Overall performance coefficient (0 to 100)
07	pc_first_semester	Numeric	Yes	Performance coefficient in the first semester of studies (0 to 100)
08	course	Categoric	No	Student course (32 different courses)

09	course_quality	Numeric	Yes	Brazilian Ministry of Education course quality (1 to 5)
10	admission	Categoric	No	Admission (ENEM/SISU, Entrance exam, Transfer, Obtaining a new title)
11	age_joining_course	Numeric	Yes	Student age when joining the course (17 to 70)
12	course_type	Categoric	Yes	Type of course (Bachelor degree, Graduation, Technologist)
13	native_campus_city	Categoric	Yes	Native from campus city (No, Yes)
14	perc_failure_first_semester	Numeric	Yes	Percentage of failure in the first semester of studies (0 to 100)
15	perc_income_minimum	Numeric	Yes	Percentage of course students who have a per capita family income of up to one minimum wage (0 to 100)
16	scholarship	Categoric	Yes	Scholarship (No, Yes)
17	applicant_course_ratio	Numeric	Yes	Applicant/Course Vacancy Ratio
18	gender	Categoric	No	Male, Female
19	time_hs_grad	Numeric	Yes	Time between finishing high school and entering the course (in years)
20	study_shift	Categoric	No	Daytime, Full-time, Nighttime
21	status_student_course	Categoric	No	*Target class.* Dropout, Graduated, Enrolled

The dataset analyzed in this study has 12,657 instances of students at the institution, of which 5,915 are labeled as *Dropout*, 2,096 as *Graduated* and 4,646 as *Enrolled*. The students labeled as *Graduated* are those who have already finished their course and ended their relationship with the institution.

The fourth stage (*Modeling*) was divided into two phases with the objective of building, testing and evaluating different models to identify the strategy that provides the best indicators regarding the prediction of dropouts. In the first phase, three models related to the Feature Selection (FS) technique were built: with no Feature Selection, with Boruta and with Correlation Coefficient.

Kursa & Rudnicki (2010) state that Boruta is a Feature Selection technique based on Random Forest capable of calculating a numerical estimate of the importance of each analyzed attribute, enabling the understanding of which is the best set of attributes capable

of providing good results in classification terms. The expectation is that these results will be better than those obtained when using all attributes of the dataset. The difference between Boruta and other techniques is that, from the initial dataset, a set of attributes (called shadow features) is artificially constructed whose importance values are calculated and compared with the values obtained by the original attributes, allowing the identifying those who really matter. In order for the results to have statistical significance, Boruta repeats this process of building the dataset several times with the shadow features, which are independent in each iteration.

Correlation Coefficient is a Feature Selection technique that calculates the dependency between attributes and, eventually, decides to remove one of them from the dataset. Such a measure is justified by the fact that when two attributes are mutually dependent, it is very likely that their occurrence and variation will be the same. In the case of classification tasks, the maintenance of only one of the correlated attributes is able to maintain, at least, the same results obtained with both. However, it is common to observe improvements in model accuracy when redundant attributes are removed (Hsu & Hsieh, 2010).

Only instances of students with *Dropout* or *Graduated* status were analyzed at this stage. Figure 2 presents the set of activities that make up the first phase of the Modeling stage.

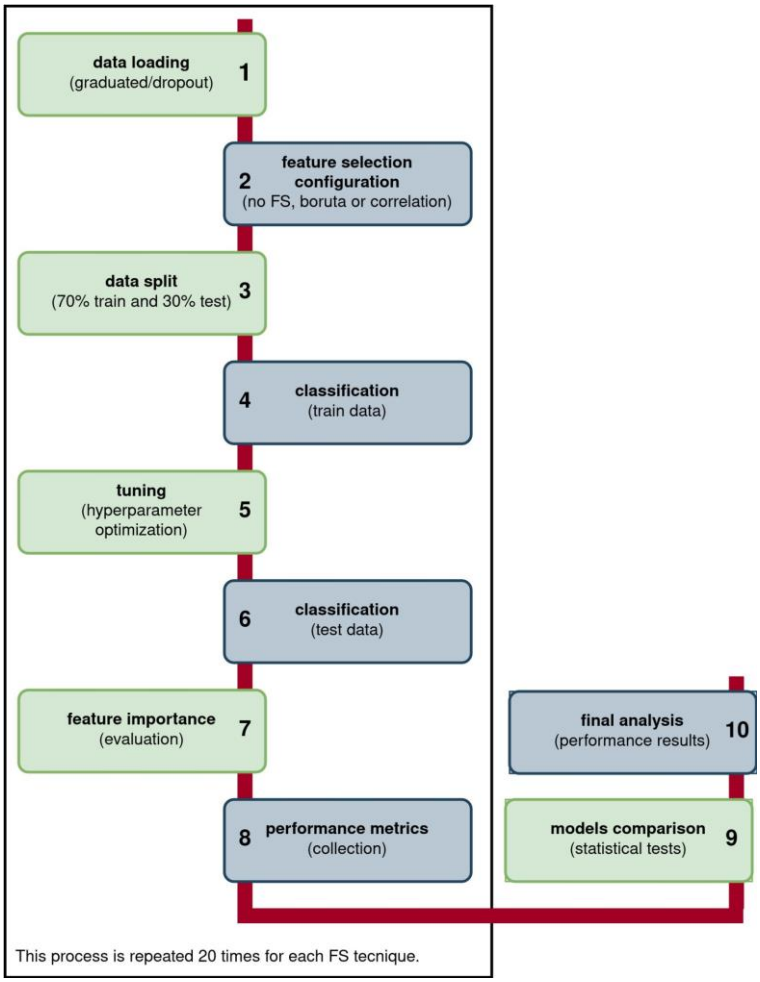

Figure 2. Activities performed in Phase 1 of the Modeling stage

The initial activity consists of loading the data to the experimental environment, followed by the configuration of the Feature Selection technique, considering that a model was created with no FS technique, another with the application of the Correlation Coefficient technique and another with the Boruta technique. In the third activity, the loaded data were separated into two independent sets to be used in the training (with 70% of the instances) and in the test (with 30% of the instances) of the classifier, maintaining the balance in relation to the target class. The fourth activity consists of the execution of the classification algorithm on the training dataset and the fifth activity seeks to optimize the hyperparameters of the model trained with the Grid Search technique. In the sixth activity, the optimized model is applied to the test dataset (with instances not used in training). Four metrics are collected in activity 6 to illustrate the performance of each algorithm: accuracy, precision, recall and Macro-F1. Activities 1 to 7 are performed repeatedly, twenty times, for each Feature Selection technique applied in this study; at the end, the average of the executions is calculated for each metric in a context, thus allowing statistical comparisons between the results. The application of ANOVA and Tukey's tests allows the statistical validation of the differences found between the results, as they are approaches that allow multiple comparisons between the various indicated groupings (Abdi & Williams, 2010).

The second phase of the Modeling stage aimed to improve the performance of the models through the use of the semi-supervised learning strategy. Instances available in the dataset labeled *Enrolled* were evaluated by the classifier, i.e., students who have a bond with the institution. Those classified as *Graduated* were then aggregated in the creation of a new prediction model based on a dataset enlarged with pseudo-labeled instances for a more balanced dataset. The activities of this phase are listed in Figure 3.

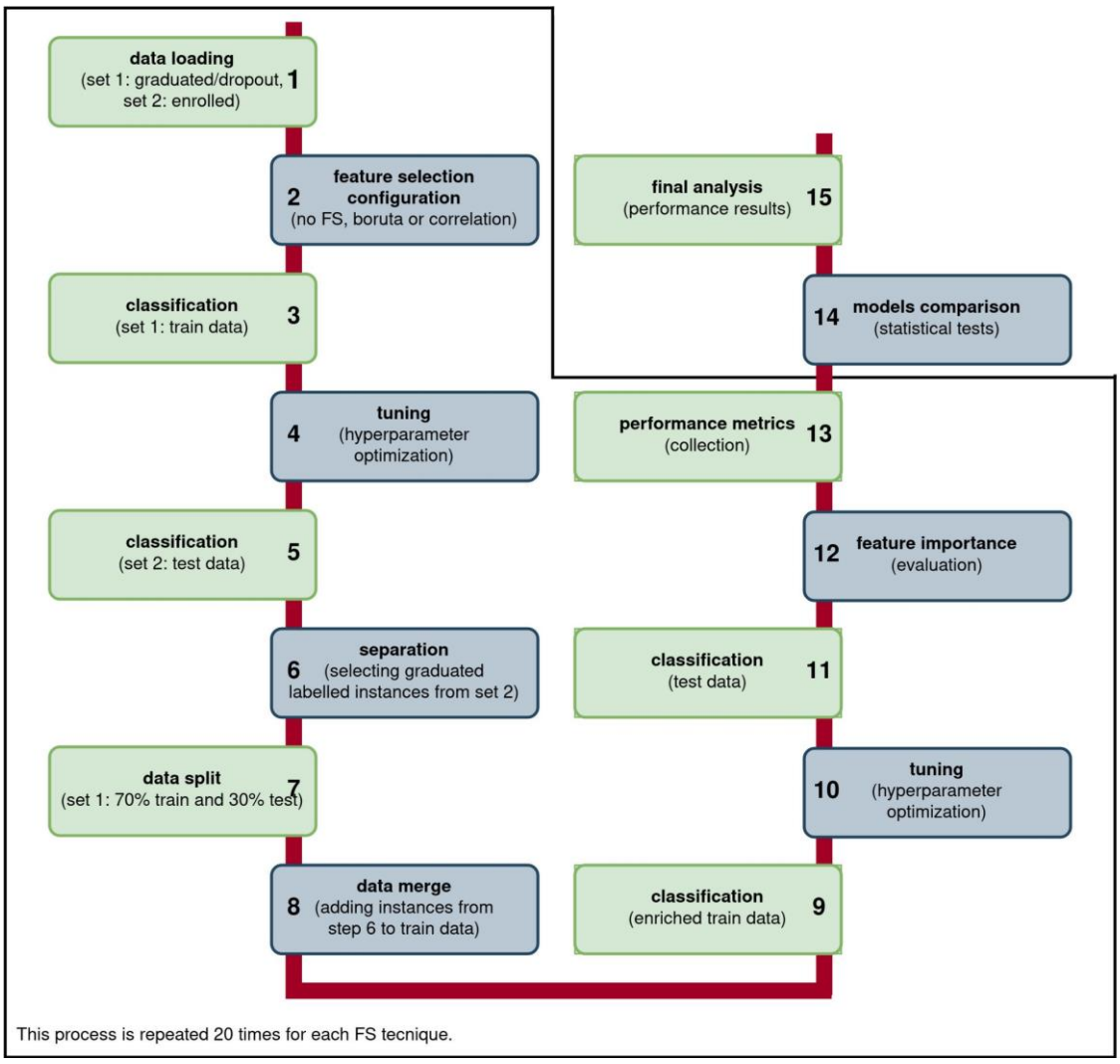

Figure 3. Activities performed in Phase 2 of the Modeling stage

Initially, two datasets are loaded, one containing instances classifieds as *Dropout* or *Graduated* (the same data as in Phase 1) and the other with only the instances labeled as *Enrolled*. Activity 2 provides for the configuration of the Feature Selection technique to be used in the construction of the model, and three possibilities were chosen, as in Phase 1: no feature selection, Correlation Coefficient or Boruta. In activities 3 and 4, model training and algorithm hyperparameters optimization occur, respectively, using the dataset composed of all instances labeled *Dropout* and *Graduated*.

The semi-supervised learning strategy becomes clearer from activity 5 onwards, when the dataset with enrolled students is analyzed by the model resulting from activity 4, aiming to identify labels such as *Graduated* that will serve for another classification process. In activity 6 such instances are separated, while in activity 7 the original dataset (with *Dropouts* and *Graduates*) is separated into training (70%) and test (30%). The instances previously classified as *Graduated* are included in the training dataset so that, in sequence, they are used in a new process of classification and tuning (activities 8, 9 and 10). The prediction on previously unknown data, i.e., from the test dataset, is carried out in activity 11. Once the model is created, the metrics are stored for future analysis within the scope of activity 12. The set of activities from 1 to 12 is repeated twenty times

for each of the three defined Feature Selection techniques, allowing a statistical comparison to occur between the resulting models (activity 13) and a final analysis of the process (activity 14).

For the creation of all the proposed models, the Random Forest (RF) algorithm was used, given its positive history of application in contexts of prediction of school dropout with Machine Learning resources, ease of interpretation and parameter optimization. Moreover, preliminary results indicated RF outperforms other classifiers such as Naive Bayes and Support Vector Machine in this study. Considering that the dataset was unbalanced in relation to the number of instances labeled *Dropout* and *Graduated*, it was decided to compare performances using the Macro-F1 metric. It is the harmonic average between the results of precision and recall, relativizing the contribution of each of these metrics to the final result, reducing the possibility of specific variations of both and well portraying the quality of the model.

For the fifth stage (*Evaluation*) of the CRISP-DM, the proposed models were evaluated, allowing adjustments and validations to be made. At the end, the performance of the models was compared based on the ANOVA and Tukey tests to identify any statistically significant differences between the results collected from the Macro-F1 metric.

The last stage foreseen by CRISP-DM (*Deployment*) was not part of the scope of this study, being a suggestion for the institution after analysis of the adequacy of the final results achieved.

4. Results and discussion

Of the institution's 18 campuses, 15 had data participating in the study. Data from one Campus were not available and the courses offered at the other two campuses do not meet the criteria defined for this study (offer undergraduate courses). Table 2 indicates the number of instances per Campus and according to their final status.

Table 2. Instances by Campus and initial label

Campus	Dropout	Graduated	Enrolled	Sum	Average dropout
A	111	0	92	203	55%
B	981	378	825	2,184	45%
C	246	3	257	506	49%
D	441	105	364	910	49%
E	768	316	354	1,438	53%
F	598	191	385	1,174	51%

G	28	0	83	111	25%
H	64	9	122	195	33%
I	627	186	414	1,227	51%
J	618	200	332	1,150	54%
K	112	24	141	277	41%
L	309	163	272	744	42%
M	314	143	273	730	43%
N	318	183	304	805	40%
O	380	195	428	1,003	38%
Total	**5,915**	**2,096**	**4,646**	**12,657**	**42%**

Regarding the results of the computational experiments, the objective of *Phase 1* (Modeling stage) was to build prediction models with and without the application of Feature Selection techniques. In this sense, models were created using the Random Forest algorithm in the unbalanced dataset. Two Feature Selection techniques were applied, Boruta and Correlation Coefficient.

To proceed with the application of the Correlation Coefficient technique, a Correlation Matrix was generated (Figure 4) to enable the understanding of which attributes were linked to each other. It was found that the absolute correlation between some attributes was significant, such as *pc_first_semester* and *perc_failure_first_semester* (-0.9), *course_type* and *course_workload* (-0.9), *time_hs_grad* and *age_joining_course* (0.9), *pc* and *perc_failure_first_semester* (-0.8), *pc* and *pc_first_semester* (0.8), *perc_income_minimum* and *campus* (-0.6) and *course_quality* and *applicant_course_ratio* (0.5). In these cases, the attribute that remained in the dataset used to create the models was the one that had the highest correlation with the target class. For example, there is a high correlation between the attributes *pc*, *pc_first_semester* and *perc_failure_first_semester*, but the attribute *pc* is the one most strongly correlated with the attribute *status_student_course* (0.6), thus being kept in the dataset.

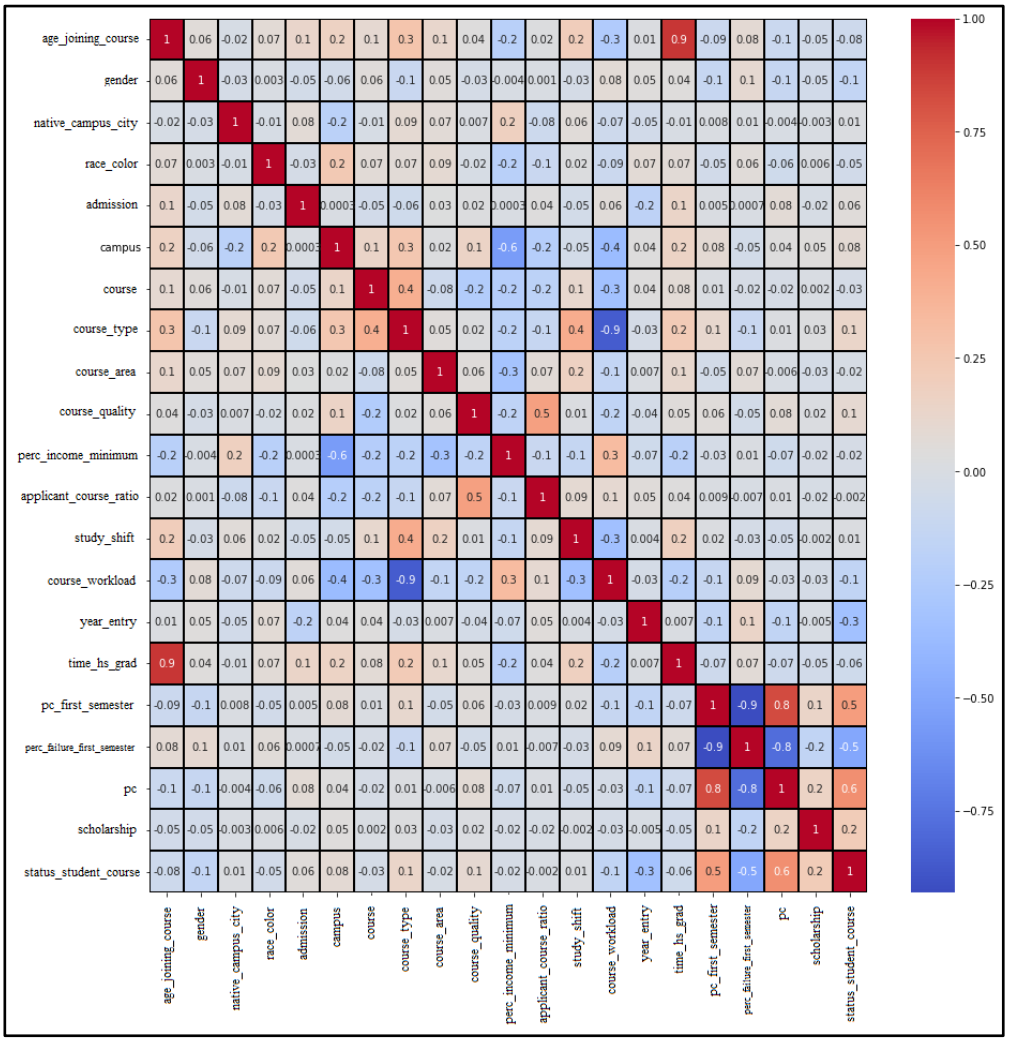

Figure 4. Correlation Matrix

The results obtained in each of the contexts are shown in Table 3, where it is possible to notice that the Correlation Coefficient technique delivered the best accuracy and Macro-F1.

Table 3. Feature Selection contexts metrics (Phase 1)

Context	Accuracy	Recall	Precision	Macro-F1
Without Feature Selection	0.8880	0.9261	0.7309	0.8155
Boruta	0.8977	0.9047	0.7570	0.8222
Correlation Coefficient	0.8988	0.9254	0.7471	0.8260

The ANOVA test was applied to the sets of results collected from the Macro-F1 metric, and it was possible to identify that the difference between the contexts was not statistically significant (F=2.16 and F critical=3.15). The Tukey Test comparing the means of the contexts confirmed this lack of significance, as seen in Table 4.

Table 4. Statistical validation – difference between contexts (Phase 1)

Comparisons	p	Significant
Without Feature Selection x Boruta	0.3945	No
Without Feature Selection x Correlation Coefficient	0.1094	No
Boruta x Correlation Coefficient	0.7414	No

It is worth mentioning that the application of the Feature Selection techniques did not improve the prediction results of the models, as it showed a similar performance to the dataset with no selected attributes.

The proposal of *Phase 2* (Modeling stage) was to use the 4,646 instances of the original dataset labeled as *Enrolled*, i. e. students who have not yet ended their bond with the institution, either due to dropout or graduation. This semi-supervised learning approach allowed those unlabeled instances to be classified by a model and, later, those identified as *Graduated* became part of the train dataset for the prediction of dropout. This action increased the number of instances labeled *Graduated* and, consequently, reduced the original imbalance of the dataset (5,915 *Dropout* and 2,096 *Graduated*).

The results obtained with the application of the semi-supervised learning strategy outperformed those from Phase 1 when analyzing the Macro-F1 metric. The accuracy remained at the same level, ~90% of correct classifications. Table 5 shows the final averages of the metrics collected in the three contexts. It is observed that the Macro-F1 of the context No Feature Selection is the best, followed by the context with Boruta and, finally, the context with Correlation Coefficient.

Table 5. Metrics of the semi-supervised contexts (Phase 2)

Context/Technique	Accuracy	Recall	Precision	Macro-F1
No Feature Selection	0.8979	0.9501	0.7944	0.8642
Boruta	0.9133	0.9104	0.8140	0.8585
Correlation Coefficient	0.8959	0.9237	0.7871	0.8476

The ANOVA test was applied to these sets and identified a statistically significant difference between the contexts (F=9.36 and F critical=3.15). Then, the Tukey test was performed comparing the means and confirmation of this statistical significance was obtained when the contexts No Feature Selection and Boruta were analyzed with the context of the Correlation Coefficient (Table 6). In other words, the results of the first two contexts are statistically equivalent and, at the same time, are better than those obtained by the third context.

Table 6. Statistical validation – difference between contexts (Phase 2)

Comparisons	**p**	**Significant**
Without Feature Selection x Boruta	0.3179	No
Without Feature Selection x Correlation Coefficient	< 0.001	Yes
Boruta x Correlation Coefficient	0.01889	Yes

It is important to highlight the increase achieved with this strategy, going from the best performance of the Macro-F1 metric in Phase 1 with a final average of 0.8260 to 0.8642 in Phase 2. The ANOVA and Tukey tests identified significant variance between the sets of results that generated such averages, with $p < 0.001$, confirming that the classification model created with the support of the semi-supervised learning strategy is more effective in improving the average Macro-F1 metric.

As explained in the Methods section, one of the activities carried out in this study was the optimization of the parameters of the built classification models. Table 7 provides the set of hyperparameters (and their respective values) for each of the three analyzed contexts. It is noteworthy that the models that used the Boruta Feature Selection technique required only twenty trees within the estimator, in contrast to the 90 with No Feature Selection and 110 with the Correlation Coefficient. The limit height of tree growth in the forest is similar between contexts, unlike the parameter that indicates the minimum number of samples that an internal node needs to have in order to be divided into other nodes and the parameter that specifies the minimum number of samples that a node must have after its division.

Table 7. Configuration of hyperparameters in each context after tuning

Parameters	**No Feature Selection**	**Boruta**	**Correlation Coefficient**
n_estimators	90	20	110
max_depth	11	9	11

max_features	1.0	1.0	sqrt
min_samples_leaf	6	2	2
min_samples_split	5	10	10

In order to understand which attributes are most related to the dropout prediction process, feature importance was analyzed for the context that presented the best performance in terms of the Macro-F1 metric, i. e. No Feature Selection. Figure 5 contains the attributes that had at least 2% importance in that model. It is noted that the *pc* attribute accounts for more than half of the general importance, while a greater balance prevails among the others. It should be noted that the two most important attributes are related to student achievement in grades, the first being the global performance coefficient and the second being the performance coefficient for the first semester only. Of the eight most important attributes, five were created specifically for this study, i. e. they were not available in the original dataset.

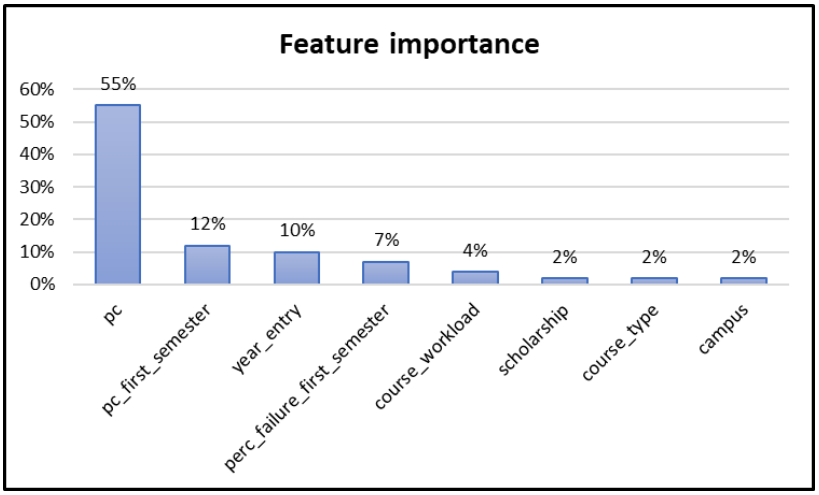

Figure 5. Feature importance for prediction

Table 8 was prepared with the purpose of providing a comparative analysis between the main characteristics and results of some literature related work to the one presented in this study. The first three columns report the author(s) of the study, Machine Learning algorithms evaluated by the proposal and which of them showed the best performance, respectively. The following columns report if hyperparameter optimization was applied, how many instances compose the dataset and if the target classes are balanced. Finally, accuracy and Macro-F1 are presented as comparative metrics.

Table 8. Comparison of results with related works

Author(s)	Algorithms evaluated	Best Algorithm	Optimized	Instances	Balanced	Accuracy	Macro-F1
Assis (2017)	C 5.0, Naive Bayes, Neural Networks and CART	CART	Yes	23,692	Yes	0.92	0.53
Berka & Marek (2021)	Random Forest, CART and Logistic Regression	Random Forest	No	3,339	No	0.91	NA
Flores, Heras & Julián (2022)	Random Forest, Random Tree, J48, OneR, Bayes Net and Naive Bayes	Random Forest	No	4,365	Yes	0.97	NA
Gonçalves, Silva & Cortes (2018)	Naive Bayes, SVM and J48	J48	No	574	No	0.98	NA
Kantorski et al. (2016)	CART, J48 and Naive Bayes	CART	No	791	No	0.73	0.46
Melo (2016)	Random Forest	Random Forest	No	32,342	Yes	0.90	0.80
Perez, Castelhanos & Correal (2018)	Random Forest, Logistic Regression and Naive Bayes	Random Forest	Yes	802	No	0.94	NA
Soares et al. (2020)	Random Forest	Random Forest	Yes	2,319	No	0.98	0.88
Teodoro & Kappel (2020)	Random Forest, Naive Bayes, KNN, Decision Trees and Neural Networks	Random Forest	Yes	376,746	Yes	0.80	0.79
This study	**Random Forest**	**Random Forest**	**Yes**	**8,011**	**No**	**0.90**	**0.86**

NA: Not available

Most of the listed studies chose to analyze at least three ML algorithms, however Random Forest was the technique that provided the best result in seven of the ten works. In this way, the choice in this study to conduct the experiments in Phases 1 and 2 using this technique not only based on the metrics collected and statistically validated, but also based on a basis coming from other publications that confirmed Random Forest as an adequate strategy to solve problems of classification of dropout students. Regarding accuracy and Macro-F1 metrics, no pattern could be seen considering hyperparameter optimization and dataset imbalance. Therefore, it is not possible to say that there is a direct relationship between these two characteristics and the increase in metrics, leading to an understanding that, no matter how similar the strategies, techniques and resources used, each study is particular enough to produce their own evidence within the broader context of Machine Learning.

5. Conclusions

School dropout can still be considered as one of the major educational problems faced by Brazilian institutions, especially those that offer undergraduate courses. In the case of public institutions, this is a problem that should involve the whole society, as society indirectly affords for the studies of individuals who may, eventually, terminate the link with the course and lose such an investment. Not a detriment exclusively to the institution, but also to the student who had an expectation when entering the course and, for various reasons, was unable to complete it. School dropout is not a recent phenomenon, but it is important that it is always on the agenda in academic discussions, so that new actions and plans are carried out to try to reduce it to acceptable levels, making student dropouts not occur due to lack of reasonable conditions for them to finish their courses.

In this context, the present study aimed to integrate Information Technology resources, especially in the area of ML, to better understand how the semi-supervised learning approach can help to increase the results of dropout prediction in a Brazilian federal educational institution. It is expected that with a broader knowledge about the public involved and its particularities, academic managers can make better decisions in the sense of not acting only after the dropout occurs, but mainly that actions are developed so that the student is not impelled to leave his course.

Regarding the ML techniques applied, the increase in the final result of the Macro-F1 metric is highlighted when comparing phases 1 and 2 proposed in the study methodology. The creation of models based on the supervised learning approach generated 82% of hits in this metric, while the semi-supervised models contributed to this indicator reaching 86%, keeping the same characteristics in terms of the dataset, Feature Selection and Grid Search. Contrary to expectations, the application of Feature Selection techniques did not bring significant improvements in the prediction capacity of the created models. One of the possibilities for this fact is that in the analyzed dataset, few attributes have great predictive capacity: three attributes represent almost 90% of the total importance.

Observing other studies related to this work, note the advance brought by the introduction of the semi-supervised learning approach as a factor capable of helping to improve the predictions about school dropout, given that the use of the supervised learning approach is still the most common in studies on this topic. This advance is confirmed as the result of Macro-F1 was the second best among related works that

presented this metric. We understand that this indicator is more suitable for the context of school dropout, compared to only the analysis of accuracy, as it involves two metrics in its composition. Furthermore, it was possible to confirm the adherence of CRISP-DM as a methodology capable of properly organizing the data mining process, which was also used by Berka and Marek (2021), Assis (2017) and Kantorski et al. (2016). As presented by Assis (2017), the techniques for balancing the dataset did not improve the collected metrics. There was also adherence to the studies by Melo (2016) and Gonçalvez, Silva and Cortes (2018) regarding the use of Feature Selection techniques.

The main objective of this study was fulfilled by presenting as a result a model for predicting school dropout with the ability to correctly classify ~90% of their predictions. Furthermore, when we analyze the Macro-F1 metric obtained in this study in comparison with baseline studies, the performance is very close to the highest found (0.86 vs. 0.88). It is precisely this final product that can be an ally of the institution's academic management in terms of having a more proactive stance on dropout, so that students are accompanied before they reach a point of no return, i. e. when the conditions offered for their continuity in the course are no longer satisfactory and dropout will occur. From the moment that the coordinator and the professors of a course have an indication of students who can potentially dropout, it is fully possible to outline actions and goals that are coherent with the profile of each one of them so that, if it occurs, the dropout is by a reason outside the scope of the institution. This type of action closer to the student tends to collaborate so that dropout levels are reduced and a relationship of trust and mutual support is created between the people involved.

The period covered by this study involved students enrolled between 2013 and 2019, especially because the COVID-19 pandemic caused the interruption of activities on some campuses and the adoption of a decentralized academic calendar. This choice can be understood both as a limitation of the present work and as an opportunity for new research, which can have their results compared with those presented here, in order to verify if the evasion that occurred after the beginning of the pandemic is caused by different factors of those surveyed in the classes that entered until 2019.

References

Abdi, H. and Williams, L. J. (2010). Newman-Keuls test and Tukey test. In *Encyclopedia of research design*, 2, 1-11. https://personal.utdallas.edu/~Herve/abdi-NewmanKeuls2010-pretty.pdf

Agrusti, F., Bonavolontà, G. and Mezzini, M. (2019). University Dropout Prediction through Educational Data Mining Techniques: A Systematic Review. In *Journal of E-Learning and Knowledge Society*, 15(3), 161-182. https://doi.org/10.20368/1971-8829/1135017

Andifes. (2019). *V Pesquisa do Perfil Socioeconômico e Cultural dos Estudantes de Graduação das Instituições Federais de Ensino Superior Brasileiras. Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior (ANDIFES)*. Retrieved May 1, 2022. https://www.andifes.org.br/wp-content/uploads/2019/05/V-Pesquisa-Nacional-de-Perfil-Socioeconomico-e-Cultural-dos-as-Graduandos-as-das-IFES-2018.pdf

Assis, L. R. S. (2017). *Perfil de evasão no ensino superior brasileiro: uma abordagem de mineração de dados.* Retrieved April 22, 2022. https://repositorio.unb.br/handle/10482/32139

Ayodele, T. O. (2010). Types of machine learning algorithms. In *New advances in machine learning*, 3, 19-48. https://doi.org/10.5772/9385

Baggi, C. A. S. and Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. In *Revista da Avaliação da Educação Superior (Campinas)*, 16(2), 355-374. https://doi.org/10.1590/S1414-40772011000200007

Balkis, M. (2018). Academic motivation and intention to school dropout: the mediation role of academic achievement and absenteeism. In *Asia Pacific Journal of Education*, 38(2), 257-270. https://doi.org/10.1080/02188791.2018.1460258

Batista, G. E. (2003). *Pré-processamento de Dados em Aprendizado de Máquina supervisionado.* Retrieved November 10, 2021. https://doi.org/10.11606/T.55.2003.tde-06102003-160219

Berka, P. and Marek, L. (2021). Bachelor's degree student dropouts: Who tend to stay and who tend to leave?. In *Studies in Educational Evaluation*, 70. https://doi.org/10.1016/j.stueduc.2021.100999

Biazus, C. A. (2004). *Sistema de fatores que influenciam o aluno a evadir se dos cursos de graduação na UFSM e na UFSC: um estudo no curso de ciências contábeis.* Retrieved January 23, 2022. https://oatd.org/oatd/record?record=oai\:repositorio.ufsc.br\:123456789\%2F87138

Brasil. (2007). *Decreto Nº 6.096, de 24 de abril de 2007.* Institui o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais - REUNI. Retrieved January 7, 2022. http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2007/Decreto/D6096.htm

Castro, A. K. S. and Teixeira, M. A. P. (2013). A evasão em um curso de psicologia: uma análise qualitativa. In *Psicologia em Estudo*, 18(2), 199-209. https://www.scielo.br/j/pe/a/YRM6f59TQX5stXQhyNK459c/?lang=pt

Ceratti, M. R. N. (2008). *Evasão escolar: causas e consequências.* Retrieved February 12, 2022. https://docplayer.com.br/19255605-Evasao-escolar-causas-e-consequencias.html

Davok, D. F. and Bernard, R. P. (2016). Avaliação dos índices de evasão nos cursos de graduação da Universidade do Estado de Santa Catarina-UDESC. In *Revista da Avaliação da Educação Superior*, 21(2), 503-522. https://doi.org/10.1590/S1414-40772016000200010

Demeter, E., Dorodchi, M., Al-Hossami, E., Benedict, A., Walker, L. and Smail, J. (2022). Predicting first-time-in-college students' degree completion outcomes. In *Higher Education*, 1-21. https://doi.org/10.1007/s10734-021-00790-9

Diniz, R. V. and Goergen, P. L. (2019). Educação Superior no Brasil: panorama da contemporaneidade. In *Revista da Avaliação da Educação Superior*, 24(3), 573-593. https://doi.org/10.1590/s1414-40772019000300002

Fialho, M. G. D. and Prestes, E. M. T. (2014). Evasão escolar no curso de pedagogia da UFPB: na compreensão dos gestores educacionais. In *Mpgoa*, 3(1), 42-63. https://periodicos3.ufpb.br/index.php/mpgoa/article/view/19005

Fisher, D. H., Pazzani, M. J. and Langley, P. (2014). *Concept formation: Knowledge and experience in unsupervised learning.* Morgan Kaufmann.

Flores, V., Heras, S. and Julian, V. (2022). Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education. In *Electronics*, 11(3), 457. https://doi.org/10.3390/electronics11030457

Gibson, B. R., Rogers, T. T. and ZHU, X. (2013). Human semi-supervised learning. In *Topics in cognitive science*, 5(1), 132-172. https://doi.org/10.1111/tops.12010

Gonçalves, T. C., Silva, J. C. and Cortes, O. A. C. (2018). Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do Instituto Federal do Maranhão. In *Revista Brasileira de Computação Aplicada*, 10(3), 11-20. https://doi.org/10.5335/rbca.v10i3.8427

Guo, J., Wang, Q. and Li, Y. (2021). Semi-supervised learning based on convolutional neural network and uncertainty filter for façade defects classification. In *Computer-Aided Civil and Infrastructure Engineering*, 36(3), 302-317. https://doi.org/10.1111/mice.12632

Hegde, V. and Prageeth, P. P. (2018). Higher education student dropout prediction and analysis through educational data mining. In *2nd International Conference on Inventive Systems and Control (ICISC)*, 694-699. https://doi.org/10.1109/ICISC.2018.8398887

Helm, J., Swiergosz, A., Haeberle, H., Karnuta, J., Schaffer, J., Krebs, V., Spitzer, A. and Ramkumar, P. (2020). Machine learning and artificial intelligence: definitions, applications, and future directions. In *Current reviews in musculoskeletal medicine*, 13(1), 69-76. https://doi.org/10.1007/s12178-020-09600-8

Hsu, H. H. and Hsieh, C. (2010). Feature Selection via Correlation Coefficient Clustering. In *Journal of Software*, 5(12), 1371-1377. https://doi.org/10.4304/jsw.5.12.1371-1377

Inep. (2019). *Resumo técnico do Censo da Educação Superior 2019*. Retrieved December 12, 2021. https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/resumo_tecnico_censo_da_educacao_superior_2019.pdf

Jagodics, B. and Szabó, E. (2022). Student burnout in higher education: A demand-resource model approach. In *Trends in Psychology*, 1-20. https://doi.org/10.1007/s43076-021-00137-4

Jia, P. and Maloney, T. (2015). Using predictive modelling to identify students at risk of poor university outcomes. In *Higher Education*, 70(1), 127-149. https://doi.org/10.1007/s10734-014-9829-7

John, T. J., Walsh, M., Raczek, A., Vuilleumier, C., Foley, C., Heberle, A., Sibley, E. and Dearing, E. (2018). The long-term impact of systemic student support in elementary school: Reducing high school dropout. In *AERA Open*, 4(4). https://doi.org/10.1177/2332858418799085

Jordan, M. I. and Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. In *Science*, 349(6245), 255-260. https://doi.org/10.1126/science.aaa8415

José, A. R., Broilo, C. L. and Andreoli, G. S. *A evasão na Unipampa – diagnosticando processos, acompanhando trajetórias e itinerários de formação.* Retrieved January 20, 2022. https://sites.unipampa.edu.br/formacao/files/2010/07/relatorio_final_evasao-na-unipampa_out20111.pdf

Kantorski, G., Flores, E., Schmitt, J., Hoffmann, I. and Barbosa, F. (2016). Predição da evasão em cursos de graduação em instituições públicas. In *Simpósio Brasileiro de Informática na Educação-SBIE*, 27(1). http://dx.doi.org/10.5753/cbie.sbie.2016.906

Kehm, B. M., Larsen, M. R. and Sommersel, H. B. (2019). Student dropout from universities in Europe: A review of empirical literature. In *Hungarian Educational Research Journal*, 9(2), 147-164. https://doi.org/10.1556/063.9.2019.1.18

Koc, M., Zorbaz, O. and Demirtas-zorbaz, S. (2020). Has the ship sailed? The causes and consequences of school dropout from an ecological viewpoint. In *Social Psychology of Education*, 23(5), 1149-1171. https://doi.org/10.1007/s11218-020-09568-w

Kursa, M. B. and Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. In *Journal of Statistical Software*, 36(11), 1–13. https://doi.org/10.18637/jss.v036.i11

Lee, S. and Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. In *Applied Sciences*, 9(15), 3093. https://doi.org/10.3390/app9153093

Martins, C. B. N. (2007). Evasão de alunos nos cursos de graduação em uma instituição de ensino superior. Retrieved April 3, 2022. https://www.fpl.edu.br/2018/media/pdfs/mestrado/dissertacoes_2007/dissertacao_cleidis_beatriz_nogueira_martins_2007.pdf

Melo, A. S. C. (2016). *Previsão automática de evasão estudantil: um estudo de caso na UFCG*. Retrieved January 15, 2022. http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/800

Mitchell, T. (1997). *Machine Learning.* McGraw-Hill.

Momm, A. M. P. and Momm, S. F. (2020). *A evasão escolar no curso superior de tecnologia em Jaraguá do Sul*. Retrieved November 12, 2021. https://repositorio.ifsc.edu.br/handle/123456789/1417

Monard, M. C. and Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. In *Sistemas inteligentes - Fundamentos e aplicações*, 1(1). https://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf

Morais, J. I., Abonizio, H. Q., Tavares, G. M., da Fonseca, A. A., and Barbon Jr, S. (2020). A Multi-label Classification System to Distinguish among Fake, Satirical, Objective and Legitimate News in Brazilian Portuguese. In *ISys - Brazilian Journal of Information Systems*, 13(4), 126–149. https://doi.org/10.5753/isys.2020.833

Musso, M. F., Hernández, C. F. R. and Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: a machine-learning approach. In *Higher Education*, 80(5), 875-894. https://doi.org/10.1007/s10734-020-00520-7

Nagai, N. P. and Cardoso, A. L. J. (2017). A evasão universitária: Uma análise além dos números. In *Revista Estudo & Debate*, 24(1). http://dx.doi.org/10.22410/issn.1983-036X.v24i1a2017.1271

Neves, C. E. B. and Martins, C. B. (2016). *Ensino superior no Brasil: uma visão abrangente*. Retrieved November 1, 2021. http://repositorio.ipea.gov.br/handle/11058/9061

Niaksu, O. (2015). CRISP data mining methodology extension for medical domain. In *Baltic Journal of Modern Computing*, 3(2), 92-109. https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/3_2_2_Niaksu.pdf

Nicoletti, M. C. (2019). Revisiting the Tinto's Theoretical Dropout Model. In *Higher Education Studies*, 9(3), 52-64. https://ideas.repec.org/a/ibn/hesjnl/v9y2019i3p52-64.html

Nonato, B. F., Nogueira, C., Lima, L. and Otoni, S. (2020). Mudanças no perfil dos estudantes da UFMG: desafios para a prática docente. In *Revista Docência do Ensino Superior*, 10, 1-21. https://doi.org/10.35699/2237-5864.2020.20463

Pascoe, M. C., Hetrick, S. E. and Parker, A. G. The impact of stress on students in secondary school and higher education. In *International Journal of Adolescence and Youth*, 25(1), 104-112. https://doi.org/10.1080/02673843.2019.1596823

Perez, B., Castellanos, C. and Correal, D. (2018). Applying data mining techniques to predict student dropout: a case study. In *Colombian Conference on Applications in Computational Intelligence (ColCACI)*, 1-6. https://doi.org/10.1109/ColCACI.2018.8484847

Rumberger, R. W. (2020). The economics of high school dropouts. In *The economics of education*, 1, 149-158. https://doi.org/10.1016/B978-0-12-815391-8.00012-4

Shirasu, M. R. and Arraes, R. A. (2016). Determinantes da evasão e repetência escolar. 2016. Retrieved March 12, 2022. https://www.anpec.org.br/encontro/2015/submissao/files_I/i12-85f3c3774c3d65741cb278e01e61db39.pdf

Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O. and Lobo, M. B. C. (2007). A evasão no ensino superior brasileiro. In *Caderno de Pesquisa*, 37(132), 641-659. https://doi.org/10.1590/S0100-15742007000300007

Soares, L. C. C., Ronzani, R., Carvalho, R. and Silva, A. (2020). Aplicação de Técnicas de Aprendizado de Máquina em um Contexto Acadêmico com Foco na Identificação dos Alunos Evadidos e não Evadidos. In *Humanidades & Inovação*, 7(8), 223-235. https://revista.unitins.br/index.php/humanidadeseinovacao/article/view/3293

Sousa, M. C. C. (2020). *Uma análise do algoritmo K-means como introdução ao aprendizado de máquinas*. Retrieved January 3, 2022. http://repositorio.uft.edu.br/handle/11612/1764

Stadler, M. J., Becker, N., Greiff, S. and Spinath, F. M. (2015). The complex route to success: complex problem-solving skills in the prediction of university success. In *Higher Education Research & Development*, 35, 1–15. https://doi.org/10.1080/07294360.2015.1087387

Teodoro, L. A. and Kappel, M. A. A. (2020). Aplicação de Técnicas de Aprendizado de Máquina para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil. In *Revista Brasileira de Informática na Educação*, 28, 838-863. http://dx.doi.org/10.5753/rbie.2020.28.0.838

Van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. In *Mach Learn*, 109, 373–440. https://doi.org/10.1007/s10994-019-05855-6

Wang, Z. and Taylor, M. E. (2017). Improving Reinforcement Learning with Confidence-Based Demonstrations. In *International Joint Conference on Artificial Intelligence (IJCAI-17)*, 3027-3033. https://doi.org/10.24963/ijcai.2017/422

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. 29-39. http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Zhu, X. and Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. In *Synthesis lectures on artificial intelligence and machine learning*, 3(1), 1-130. https://doi.org/10.2200/S00196ED1V01Y200906AIM006