Socioeconomic Data Mining and Student Dropout: Analyzing a Higher Education Course in Brazil

Oldair Luiz Gonçalves
Dept. of Education, Federal Institute of Espírito Santo (IFES)
Espírito Santo, Brazil

Walber Antonio Ramos Beltrame
Dept. of Education, Federal Institute of Espírito Santo (IFES)
Espírito Santo, Brazil

Abstract

This paper aims to analyze the student dropout from a higher education course, in the city of Guarapari, Espírito Santo, Brazil, through the use of the computational tool known as data mining. The objective was to investigate the possible scenarios for the early identification of students with higher risk of dropping out by analyzing socioeconomic data from business school graduates between 2014 and 2018 with the use of information extracted from the academic system. The methodology used was the experimental research, from a quantitative approach through a comparative analysis of data resulting from the processing of computational algorithms. After the analysis, it was concluded that computational techniques can be used to help administrators to plan pedagogical and administrative actions and that the combination of socioeconomic data with school performance information, using the tool, can yield advantageous results, allowing the fight against evasion to be seen as an early and continuous practice.

Keywords: Academic System; Data Mining; Student Dropout.

1. Introduction

The topic of students' dropout in higher education has great relevance in the educational context, as it has a meaningful impact over the educational institutions and systems, being of extreme importance in the life path of students who choose to conclude or not an undergrad course. It is a phenomenon of complex nature, where many variables and factors interact. Due to its wide scope, this topic can be presented and interpreted through many different perspectives, allowing for a great variety of analysis. For this reason, it has been present in many different forms.

The research conducted over the topic of student dropout is relevant in the discussion of an educational institution's commitment to current and future issues which pertain to students' formation. It is observed that, in the period between the students' admission and graduation many interactions happen with the purpose of encouraging students to finish their course. On the other hand, the will to drop out of a course can also be cultivated in many ways.

The Brazilian Ministry of Education [6], through the Secretariat of Higher Education, created the

Special Commission of Studies on the Student Dropout in Brazilian Public Universities and identified three groups of factors of great relevance in higher education dropout rates. These are the external factors, with no relation to the educational institution, the internal factors, with direct relation to the educational institution and the individual factors, which are inherent to the students' life and background.

Some of the external factors observed are the job market related to the course, social recognition and depreciation of the chosen career, the economic situation, governmental policies and difficulties to keep updated with technological, economic and social contemporary changes. Meanwhile, some of the internal factors related to the educational institutions are academia inherent issues, such as lack of clarity in the chosen course's pedagogical project, low didactic and pedagogical levels, possible institutionalized culture of faculty under appreciation and a lacking support structure to higher education courses. Lastly, the individual factors, inherent to the students' life and background, can be related to personality, lack of developed study skills, previous schooling, early choice of profession, lack of information about the chosen course, difficulties due to multiple failings or poor attendance.

Therefore, studies on students' dropout comprehend a series of analysis, which regard economic, emotional, administrative and pedagogical activities [8], related to the academic community and based on institutional competences and evaluations. As a result, the definition of analytical parameters for scientific research are complex. Another factor that should be considered when analyzing dropout rates is the concept which is being used, since each institution and researcher has its favored approaches. Therefore, dropout rates can be analyzed based on the student leaving the course, the campus, the institution or the educational system altogether [8].

Aside from the previous definitions, it is also essential to note that the students' drop out can cause poorer productive efficiency in companies, motivate loss of national competitiveness and the lack of specialized workforce [22]. Furthermore, it impedes social change and the improvement of the individuals' quality of life [18], since the dropouts consider dropping out as the failure to reach previously set professional goals [5].

This research analyses the application of data mining techniques in the undergrad Business Administration major at Federal Institute of Espírito Santo (IFES), Guarapari campus. The analysis was conducted using data provided by the Academic Registry Coordination, taking into account students enrolled between 2014 and 2018, and aiming at the early identification of students with a tendency of dropping out.

Differently from what was observed on other similar papers, where the analyzed data related to scholarly performance through grades, this paper assumes the existence of other evaluation criteria, which can be used in computational techniques, such as the analysis of socioeconomic indicators. Therefore, it is estimated that the risk of a student dropping out could be calculated even before the beginning and in the early months of the school year, previously to teacher's first evaluations, increasing the preventative measures the Institute could take.

This paper is composed by the following sections: Chapter 2 presents the literature review on the subject of dropout rates, while Chapter 3 encompasses the contextualization of the study and a description of the collected data. Next, Chapter 4 brings the description of the algorithms applied in the data analysis, Chapter 5 brings forth the results obtained in the analysis and the subsequent discussion and, lastly, Chapter

6 presents the conclusion of the study and possible paths for future papers.

2. Literature Review

This chapter is divided in two main subsections. The first one presents the concepts of dropout rates in higher education, building the foundation for the paper's proposal and applied measuring criteria. Meanwhile, the second subsection contextualizes the computational tools used in the mining of educational data, which will be exemplified in the following chapters.

2.1 The student dropout

The study of student dropout rates in Brazilian public higher education is part of a discussion that takes into account the institutional commitment of a University with issues and problems of its time and the future perspective of its reality. In the time between a students' enrollment and graduation, an interaction takes place, which must be reconstructed if there is an intention of learning something about the university, the difficulties faced by its students and the questions posed to them.

The term student dropout has been object of different investigations and scientific research, with the purpose of deepening the knowledge about the motivations and impacts of the topic in the educational context. To characterize the many aspects related to the topic appropriately, this paper defines in a consolidated form the interpretation of the topic, presents selected theories that support the given concepts and, lastly, demonstrates the model used for this analysis.

In addition, student dropout in higher education can be considered as a complex educational phenomenon, defined as the interruption of the educational process and becoming, with time, a disturbing problem to public as well as private educational institutions' administrators. It can also be understood as the loss of students in many levels of education, causing, amongst other, social, academic and economic consequences [18]. Therefore, the interest in statistical methods capable of the early identification of likely candidates do drop out is constantly on the rise.

Complementing the definitions shown above, [7] classifies student dropout in three new types, which are: (1) Final dropout, when students abandon the institution in definitive, regardless of motivation; (2) Temporary dropout, when the voluntary interruption of the course occurs, for a period which can vary from one to ten semesters and; (3) Course dropout, when an internal transfer occurs and the student is relocated in a different course within the same institution.

According to these prerogatives, [22] separates student dropouts in two. The first is the annual average dropout, represented by the percentage of students enrolled in the educational system, institution or different courses who, even not having concluded their studies, have not registered in the following semester. The second is the total student dropout, which takes into account the number of students who, having enrolled in a specific course in the past, have not graduated in a set period of time, which is called graduation index.

2.1.1 Presenting some theories

The learning processes a student goes through after enrolling in a higher education course can and

must be analyzed in different ways. This occurs because, while there are those who complete theirs studies in the regularly set period of time, some need more time to complete their studies and others abandon their studies, dropping out of the course.

Those who study the subject present different analytical aspects, such as: (1) The presentation of theoretical models and the development of theories, which aim at explaining the process of dropping out of a course [25], [14], [4]; (2) The development of models which take into consideration the characteristics of the students, of the institutions and of the relationship between them, searching for indicators which allow the student dropout to be anticipated [9], [2] and; (3) Aspects that allow for a greater intervention in the educational system, aiming at the articulation of proper institutional actions which aid in the retention of the students [24], [29], [3].

The students' dropout and retention processes in higher education, which have been studied since the 1950's, are based in different perspectives. Even though the priority consistently lies in the relationship between the student and the institution and in the circumstances that occasion for the rupture of this relationship, before taking into account the school's curriculum.

Through the presentation of different theories, this literature review demonstrates that the phenomenon has accumulated a growing amount of attention by researchers, allowing for a better understanding of the different research focal points. These theories arise from different approaches, with the most noteworthy being, according to [18]: (1) The sociological approach, which follows the ideas of [28], [25], [26], [2], [13], [1]; (2) The psychological approach, which follows the ideas of authors such as [14], [9], [27], [15], [10]; (3) The study of organizational, interaction and economic factors [11], [18], [26] and; (4) The pedagogical approach [4].

2.1.2 Measuring the student dropout rate

There are different ways to measure the student dropout, all of which are based on three fundamental pillars: the concept of student dropout adopted according to the phenomenon being studied; the nature of the data available for analysis and; the calculation procedure. According to [23], these three aspects are intertwined, since the conceptual definition impacts on data collection and processing.

For this paper, the concept of student dropout taken into consideration is the same used by IFES, which considers as a dropout a student who hasn't yet conclude his or her course but has failed to register for subsequent curricular components without officially interrupting his or her course or requesting a transference. It has also been established that a student is not considered a dropout if he or she has transferred courses or campuses, as long as he or she continues to study in the Institute.

2.2 Educational Data Mining

The expression data mining refers to the use of technologies supported by computational learning algorithms, which automatize the creation of analytical models through structured data, extracting or indicating patterns, which will then be used for information grouping (descriptive modeling), classification (prescriptive modeling) or prediction (predictive modeling) [12].

In regards to the descriptive modeling, it's possible to highlight the methods for grouping of similar entries (clustering); detection of anomalies for the discovery of discrepant values; association rules to

determine the relationship between entries; statistical component analysis which point out relationships between variables and groups of data affinities [12]. In that which pertains to prescriptive modeling, it's possible to enumerate the techniques for analyzing, filtering and transforming non-structured data (such as files, tests, images and videos) for prepossessing (data preparation and exploration) and post processing (model validation, foundation and monitoring) though rules classifiers and identifiers [12]. Lastly, in regards to predictive modeling, it is important to highlight the approaches that accomplish data regression as a metric for the relationship between dependent and independent variables; neural networks to identify patters; decision trees for probable occurrences; and vector machines for prediction through machine learning [12].

In the educational field, according to [20], [21], Data Mining is applied in academic systems with the purpose of discovering potential groups of students with similar characteristics, identifying behaviors after pedagogical strategies have been employed, detecting errors or incorrect use in virtual learning environments, proposing new studies or didactic resources based on automatic evaluations and finding demotivated students with a tendency towards dropping out [16], [17].

3. Case Study

The foundation of the School for Craftsmen Apprentices of Espírito Santo through the Decree number 7,566 from September 23rd 1909 by then President Nilo Peçanha started the official offer of vocational courses on a Federal level. With initial focus on free primary and professional education, it had as its main goal to train youths who could and wanted to learn a trade through practical education and the transfer of practical knowledge needed to act as factory workers and foremen [19].

After its establishment, it went through many changes, and the names of Vocational School of Vitória (ETV), Vocational Federal School of Espírito Santo (ETFES), Federal Center of Vocational Education of Espírito Santo (CEFET-ES), when it also becomes a higher education institution. In 2008, through the Law number 11,892, the Federal Institute of Espírito Santo (IFES) was established by the unification of CEFET-ES and local agrotechnical schools.

Following policies of expansion away from the metropolitan areas, so as to reach poorer communities where access to education is limited, which started in 2002, the IFES currently caters to around 22 thousand students enrolled in over 90 vocational courses, 50 undergrad courses, 20 specialization courses and 10 master programs. IFES is present in all regions of state, working out of 22 campuses and 35 hubs for distance learning.

Inaugurated in April 19th 2010, as part of the government policy for vocational schools expansion, the Guarapari campus, located South of the metropolitan area of Vitória, Espírito Santo, began activities offering a Business Vocational High School Course, with the enrollment of 42 students happening every 6 months through an open entrance exam. In 2014, following the proposal for vertical education, the undergrad Business course was founded, accessible by both students from IFES vocational high school courses and students who graduated high school in other private and public institutions.

4. Data Mining Tools and Algorithms

Data Mining is being frequently employed with both scientific and commercial purposes, due to its capability of analyzing a great amount of data in a fast and trustworthy manner, extracting relevant information through software-consolidated tools, which possess academically accepted algorithms, enhancing visualization resources and data analysis in a facilitated way with a low learning curve [12].

The aim of this paper is to use Mining Data techniques to identify the existing relationships between socioeconomic data extracted from the academic system and student dropout. With this purpose, the software Orange Mining was chosen as it is an accessible tool of open use with many versions of the main algorithms described by [12], simple a drag and drop interface and the availability of statistical resources for data analysis and performance comparison.

Eleven classification algorithms available in the software were selected: A_1 (*CN2 rule inducer*), A_2 (*kNN*), A_3 (*Tree*), A_4 (*Random Forest*), A_5 (*SVM*), A_6 (*Logistic Regression*), A_7 (*Naive Bayes*), A_8 (*AdaBoost*), A_9 (*Neural Network*), A_{10} (*SGD*) and A_{11} (*Stacking*). These algorithms have the purpose of finding mathematical models that define classes, which, in this article, relate to student dropout and no student dropout (or student retention). The obtained model identifies new examples of unknown classes based on training routines that use data labeled in the initially described classes, which can be denominated as supervised learning. This learning is verified by a set of tests applied to the model [12].

The tool also provides the identification of classes (as seen in the columns of Tables 1 and 2, in the following section) that have greater statistical influence over the predictive class (student dropout and no student dropout), regardless of the learning algorithm. The classes here identified are: Info Gain (Information gain) - expected amount of information (entropy decrease); Gain ratio - relation between Info Gain and the intrinsic information of the attribute, which decreases the tendency to multiple value characteristics; Gini - disparity between the values and frequency attribution; X^2 (*Chi2*) - dependency between the resource and the class as a quadratic statistical measure; ReliefF - an attribute's capability to distinguish classes in similar data conditions. In this manner, it is possible to establish an order of the most important classes and adopt measures to improve the algorithms, such as disposal of attributes [12].

To evaluate the test results, it is possible to adopt statistical metrics common to the computational area. Some of them are: Accuracy - proportion of correctly classified examples, not taking into account what is positive or negative; Sensibility (or Recall) - proportion of true positive results between all positive instances; Precision - proportion of true positives in all instances classified as positive; F1 - harmonic weighted average of precision and recall; and ROC - diagnostics capabilities as the discrimination threshold varies [12].

5. Data Mining Algorithm Experimentation

As a way to evaluate the training basis and the relation between the classes and the data discrimination, an evaluating resource in the order of importance of the attributes was executed and the results were presented in Table 1 (below), ordered according to the degree of importance in a decreasing manner. It is possible to see that frequency (attendance rates) is a prevalent indicator in this perspective even though the data is more closely related to academic performance (attendance and absences) than to the social economic

descriptive, thus, it is deductively associated as a predictor of student dropout. However, in situations where there is a desire to predict the likelihood of a student dropping out even before the beginning of the school year, or during the first school semester, it is not possible to evaluate this information given the fact that attendance and absences are fed into the system later on into the semester. Therefore, the experiments were divided in two data basis for training and tests. The first one considers attendance information (frequency and frequency band) while the other one does not.

The third class in this order, C_6 - City, is an important social economic data, which can be related to difficulties of transportation for students who reside in locations that are not near the Institute, a possible cause for students' dropout. In Table 7 it is possible to observe that this attribute has a high influence in dropout rates, corroborating the previous statement. However, an aspect that must be considered in this analysis is the fact that data is not always properly updated or registered in the system. A student might move to a different address after the initial enrollment period and not inform the academic registry, which partially invalidates this conclusion.

Table 1. Class evaluation through training basis

		Info Gain	Gain ratio	Gini	X^2	ReliefF
C_8	Band (frequency)	0.25	0.28	0.14	12.62	0.06
C_7	Frequency (attendance)	0.18	0.09	0.10	27.90	0.08
C_6	City	0.12	0.07	0.00	4.85	0.02
C_4	Previous Educational Institution	0.06	0.04	0.00	0.92	0.01
C_9	Color/Race	0.05	0.03	0.03	0.16	0.11
C_3	Family Income	0.04	0.02	0.03	3.36	0.00
C_{10}	Age	0.04	0.02	0.02	5.18	0.02
C_{11}	Age range	0.03	0.02	0.02	2.79	0.08
C_5	Type of Previous Educational Institution	0.00	0.03	0.00	0.00	0.00
C_1	Gender	0.00	0.00	0.00	0.50	0.04
C_2	Type of Enrollment	0.00	0.00	0.00	0.04	0.00

Source: the author

Other attributes have also suffered from lacking data, reducing the assertion in validating or not given attributes as meaningful, such as can be observe in class C_5 - Type of Previous Educational Institution. Another important point is that data such as Gender and School age, which are largely homogeneous due to the tendency of the course in attracting women of average adult age, are attributed low relevance in the prediction modeling learning algorithms.

5.1 Cross Validation Experiment

The first experiment on the database used a technique where it is divided into ten sets of cross validation and, after ten consecutive algorithm interactions, with their own optimal setting values, a result is achieved where the mean of these results for each statistical index. The results were described in the

Tables 2 and 3, in which R, A, F, P and S are the respective values for ROC, Accuracy, F1, Precision and Sensibility (Recall), in that which pertains exclusively to student dropout, exclusively to no student dropout and the mean, in a decreasing order according to ROC.

For the database that takes into consideration school attendance, the results showed Accuracy of up to 78.7% with the SVM algorithm, similar to the results of experiments that took coursework grades into consideration [17]. For the other database, without the school attendance classes, performance was inferior, with the best Accuracy reaching 66.7%. This values can still be considered satisfactory to the point where they can be used to direct pedagogical intervention initiatives aiming towards the prevention of students' dropout in initial years.

Table 2. Algorithms' results taking into consideration database with school attendance attributes

	Mean (%)					Dropout (%)					No Dropout (%)				
	R	A	F	P	S	R	A	F	P	S	R	A	F	P	S
Stacking	84.4	75.9	74.8	75.5	75.9	84.4	75.9	61.4	73.0	52.9	84.4	75.9	82.5	76.9	88.9
Random Forest	81.5	78.0	75.8	80.3	78.0	81.5	78.0	59.7	88.5	45.1	81.5	78.0	84.9	75.7	96.7
Tree	81.3	76.6	76.2	76.2	76.6	81.3	76.6	65.3	70.5	60.8	81.3	76.6	82.4	79.4	85.6
Naive Bayes	80.8	71.6	72.1	74.1	71.6	80.8	71.6	65.5	58.5	74.5	80.8	71.6	75.9	82.9	70.0
SVM	80.7	78.7	76.1	82.7	78.7	80.7	78.7	59.5	95.7	43.1	80.7	78.7	85.6	75.4	98.9
Neural Network	80.0	73.0	71.9	72.3	73.0	80.0	73.0	56.8	67.6	49.0	80.0	78.8	80.4	75.0	86.7
Logistic Regression	76.7	75.2	73.0	75.9	75.2	76.7	75.2	55.7	78.6	43.1	76.7	78.9	82.8	74.3	93.3
kNN	76.6	73.0	72.3	72.3	73.0	76.6	73.0	58.7	65.9	52.9	76.6	79.0	80.0	76.0	84.4
CN2 rule inducer	71.9	65.2	65.9	72.4	65.2	71.9	65.2	58.8	51.5	68.6	71.9	79.1	69.9	78.1	63.3
SGD	69.9	75.9	74.2	72.5	75.9	69.9	75.9	58.5	77.4	47.1	69.9	79.2	83.0	75.5	92.2
AdaBoost	63.4	66.0	66.1	72.6	66.0	63.4	66.0	53.8	52.8	54.9	63.4	79.3	73.0	73.9	72.2

Source: the author

Table 3. Algorithms' results taking into consideration database without school attendance attributes

	Mean (%)					Dropout (%)					No Dropout (%)				
	R	A	F	P	S	R	A	F	P	S	R	A	F	P	S
Naive Bayes	66.5	61.0	61.6	67.4	61.0	66.5	61.0	58.0	47.5	74.5	66.5	61.0	63.6	78.7	53.3
Logistic Regression	64.9	66.7	64.8	65.0	66.7	64.9	66.7	44.7	55.9	37.3	64.9	66.7	76.1	70.1	83.3
CN2 rule inducer	62.4	61.7	61.1	60.8	61.7	62.4	61.7	43.7	46.7	41.2	62.4	61.7	71.0	68.8	73.3
Neural Network	62.0	60.3	58.9	58.3	60.3	62.0	60.3	37.8	43.6	33.3	62.0	60.3	70.8	66.7	75.6
AdaBoost	60.9	64.5	64.2	64.0	64.5	60.9	64.5	49.0	51.1	47.1	60.9	64.5	72.8	71.3	74.4
kNN	59.9	63.8	51.0	59.0	63.8	59.9	63.8	3.8	50.0	2.0	59.9	63.8	77.7	64.0	98.9
Random Forest	59.7	63.1	59.1	59.7	63.1	59.7	63.1	31.6	48.0	23.5	59.7	63.1	74.8	66.4	85.6
SVM	59.3	65.2	54.0	68.6	65.2	59.3	65.2	10.9	75.0	5.9	59.3	65.2	78.4	65.0	98.9

SGD	58.7	66.0	63.5	63.9	66.0	58.7	66.0	41.5	54.8	33.3	58.7	66.0	76.0	69.1	84.4
Stacking	54.6	64.5	52.5	65.3	64.5	54.6	64.5	7.4	66.7	3.9	54.6	64.5	78.1	64.5	98.9
Tree	46.3	53.9	53.1	52.5	53.9	46.3	53.9	31.6	34.1	29.4	46.3	53.9	65.2	62.9	67.8

Source: the author

5.2 Randomized Process Experiment

The second experiment used a randomized process to select the test samples, with ten algorithms runs, according to the results shown in Tables 4 and 5. When comparing the results with the first experiment, it is possible to observe that the two manners of dividing the database do not significantly alter the results. This can be considered a good indicator that the attributes applied are enough to execute a prediction of students with a higher risk of dropping out during the first semester of the course.

Table 4. Algorithms' results taking into consideration database with school attendance attributes

	Mean (%)					Dropout (%)					No Dropout (%)				
	R	A	F	P	S	R	A	F	P	S	R	A	F	P	S
Stacking	80.0	76.9	76.0	76.3	76.9	80.0	76.9	61.6	71.2	54.3	80.0	76.9	83.5	78.9	88.6
SVM	79.8	79.4	76.9	82.1	79.4	79.8	79.4	58.9	92.2	43.3	79.8	79.4	86.2	76.9	98.1
Naive Bayes	79.3	72.3	72.9	75.0	72.3	79.3	72.3	64.7	57.3	74.4	79.3	72.3	77.2	84.3	71.2
Random Forest	78.8	79.4	78.2	79.4	79.4	78.8	79.4	64.0	79.3	53.7	78.8	79.4	85.5	79.4	92.7
Logistic Regression	77.8	77.5	75.9	77.5	77.5	77.8	77.5	59.4	77.5	48.2	77.8	77.5	84.4	77.5	92.7
Neural Network	75.8	73.5	72.8	72.7	73.5	75.8	73.5	57.5	63.7	52.4	75.8	73.5	80.8	77.4	84.5
kNN	75.0	72.5	71.9	71.7	72.5	75.0	72.5	56.6	61.4	52.4	75.0	72.5	79.9	77.1	82.9
Tree	72.7	71.3	70.3	70.2	71.3	72.7	71.3	53.1	60.0	47.6	72.7	71.3	79.3	75.4	83.5
SGD	69.7	75.4	74.2	74.7	75.4	69.7	75.4	57.9	69.8	49.4	69.7	75.4	82.6	77.2	88.9
AdaBoost	66.0	66.9	67.4	68.3	66.9	66.0	66.9	55.2	51.3	59.8	66.0	66.9	73.7	77.2	70.6
CN2 rule inducer	65.1	54.2	55.3	60.4	54.2	65.1	54.2	47.6	39.1	61.0	65.1	54.2	59.3	71.4	50.6

Source: the author

Table 5. Algorithms' results taking into consideration database without school attendance attributes

	Mean (%)					Dropout (%)					No Dropout (%)				
	R	A	F	P	S	R	A	F	P	S	R	A	F	P	S
Naive Bayes	67.3	59.2	60.0	66.8	59.2	67.3	59.2	54.8	44.1	72.6	67.3	59.2	62.7	78.6	52.2
Neural Network	66.8	65.6	64.6	64.2	65.6	66.8	65.6	44.4	49.6	40.2	66.8	65.6	75.1	71.8	78.8
Logistic Regression	64.4	64.8	63.4	62.9	64.8	64.4	64.8	41.5	48.0	36.6	64.4	64.8	74.8	70.7	79.4
Random Forest	61.7	64.8	61.4	61.4	64.8	61.7	64.8	33.2	47.2	25.6	61.7	64.8	76.1	68.8	85.1
kNN	61.0	65.0	52.9	53.5	65.0	61.0	65.0	3.4	30.0	1.8	61.0	65.0	78.6	65.7	97.8
Stacking	60.1	64.8	55.3	57.5	64.8	60.1	64.8	11.5	40.7	6.7	60.1	64.8	78.0	66.2	94.9

SGD	59.5	65.4	64.2	63.8	65.4	59.5	65.4	43.2	49.2	38.4	59.5	65.4	75.1	71.3	79.4
AdaBoost	54.0	57.5	57.7	57.9	57.5	54.0	57.5	38.9	38.2	39.6	54.0	57.5	67.4	68.1	66.8
CN2 rule inducer	53.7	56.5	56.4	56.3	56.5	53.7	56.5	35.7	36.0	35.4	53.7	56.5	67.1	66.8	67.4
Tree	53.3	56.2	55.6	55.1	56.2	53.3	56.2	32.3	34.2	30.5	53.3	56.2	67.7	65.9	69.6
SVM	49.5	65.8	55.6	60.9	65.8	49.5	65.8	10.9	50.0	6.1	49.5	65.8	78.9	66.5	96.8

Source: the author

5.3 Test Base Experiment

For the last experiment, the real data obtained from the Academic Registry system, considering students enrolled between 2017 and 2018, was used as the database. Even if these students are still in the first few semesters if their course and might still drop out, this experiment expects to ascertain the behavior of the algorithms in real situations and observe if the results can be used as guidelines for specific academic actions. As can be observed in Tables 6 and 7 (below) the indexes are even higher than those seen in previous experiments and those presented by [17], presenting an Accuracy of 90.5% for the database that takes into account school attendance and 86.3% for the database that does not. In this manner, it is possible to determine that social economic attributes are effective for predictive systems of Students' dropout in the Business Administration undergrad course at IFES, Guarapari campus. As a consequence, this study may be used as a research model in other institutions and undergrad courses.

Table 6. Algorithms' results taking into consideration database with school attendance attributes

	Mean(%)					Dropout(%)					NoDropout(%)				
	R	A	F	P	S	R	A	F	P	S	R	A	F	P	S
Stacking	85.5	90.5	89.9	89.7	90.5	85.5	90.5	57.1	66.7	50.0	85.5	90.5	94.7	93.0	96.4
kNN	84.7	89.5	89.1	88.8	89.5	84.7	89.5	54.5	60.0	50.0	84.7	89.5	94.0	92.9	95.2
Naive Bayes	83.2	77.9	81.0	87.7	77.9	83.2	77.9	46.2	33.3	75.0	83.2	77.9	86.1	95.6	78.3
Random Forest	82.6	88.4	87.7	87.3	88.4	82.6	88.4	47.6	55.6	41.7	82.6	88.4	93.5	91.9	95.2
Tree	81.4	89.5	88.5	88.2	89.5	81.4	89.5	50.0	62.5	41.7	81.4	89.5	94.1	92.0	96.4
SVM	78.0	91.6	90.3	91.0	91.6	78.0	91.6	55.6	83.3	41.7	78.0	91.6	95.3	92.1	98.8
CN2 rule inducer	77.9	66.3	72.0	87.3	66.3	77.9	66.3	38.5	25.0	83.3	77.9	66.3	76.8	96.4	63.9
Logistic Regression	74.0	88.4	87.7	87.3	88.4	74.0	88.4	47.6	55.6	41.7	74.0	88.4	93.5	91.9	95.2
Neural Network	69.8	73.7	77.4	84.2	73.7	69.8	73.7	35.9	25.9	58.3	69.8	73.7	83.4	92.6	75.9
SGD	69.5	71.6	76.0	85.2	71.6	69.5	71.6	37.2	25.8	66.7	69.5	71.6	81.6	93.8	72.3
AdaBoost	64.2	74.7	78.0	83.1	74.7	64.2	74.7	33.3	25.0	50.0	64.2	74.7	84.4	91.5	78.3

Source: the author

Table 7. Algorithms' results taking into consideration database without school attendance attributes

	Mean(%)					Dropout(%)					NoDropout(%)				
	R	A	F	P	S	R	A	F	P	S	R	A	F	P	S
RandomForest	72.4	82.1	83.7	86.3	82.1	72.4	82.1	45.2	36.8	58.3	72.4	82.1	89.3	93.4	85.5
Stacking	72.4	86.3	81.0	76.2	86.3	72.4	86.3	0.0	0.0	0.0	72.4	86.3	92.7	87.2	98.8
SVM	71.4	73.7	77.2	82.9	73.7	71.4	73.7	32.4	24.0	50.0	71.4	73.7	83.7	91.4	77.1
AdaBoost	64.7	69.5	74.2	83.4	69.5	64.7	69.5	32.6	22.6	58.3	64.7	69.5	80.3	92.2	71.1
SGD	64.7	69.5	74.2	83.4	69.5	64.7	69.5	32.6	22.6	58.3	64.7	69.5	80.3	92.2	71.1
NaiveBayes	61.7	46.3	54.4	78.6	46.3	61.7	46.3	21.5	13.2	58.3	61.7	46.3	59.2	88.1	44.6
LogisticRegression	61.5	65.3	70.9	81.2	65.3	61.5	65.3	26.7	18.2	50.0	61.5	65.3	77.2	90.3	67.5
NeuralNetwork	58.0	57.9	65.0	79.7	57.9	58.0	57.9	23.1	15.0	50.0	58.0	57.9	71.0	89.1	59.0
kNN	57.3	84.2	79.9	76.0	84.2	57.3	84.2	0.0	0.0	0.0	57.3	84.2	91.4	87.0	96.4
Tree	56.9	56.8	64.1	79.5	56.8	56.9	56.8	22.6	14.6	50.0	56.9	56.8	70.1	88.9	57.8
CN2ruleinducer	50.5	57.9	65.0	79.7	57.9	50.5	57.9	23.1	15.0	50.0	50.5	57.9	71.0	89.1	59.0

Source: the author

As described in Tables 1 and 2, by not taking attendance into consideration, the classification presents an elevated false positive percentage (between 40 and 50%), which implies a significant error in the prediction and, therefore, an aspect to be improved in future studies.

Table 8. Confusion Matrix taking into consideration database with school attendance attributes

	Predicted																						
	A_1		A_2		A_3		A_4		A_5		A_6		A_7		A_8		A_9		A_{10}		A_{11}		
	E	N	E	N	E	N	E	N	E	N	E	N	E	N	E	N	E	N	E	N	E	N	Actual	
E	10	2	6	6	5	7	5	7	5	7	5	7	9	3	6	6	7	5	8	4	6	6	12	E
N	30	53	4	79	3	80	4	79	1	82	4	79	18	65	18	65	22	61	23	60	3	80	83	N
	40	55	10	85	8	87	9	86	6	89	9	86	27	68	24	71	29	66	31	64	9	86	95	

Source: the author

Table 9. Confusion Matrix taking into consideration database without school attendance attributes

	Predicted																						
	A_1		A_2		A_3		A_4		A_5		A_6		A_7		A_8		A_9		A_{10}		A_{11}		
	E	N	E	N	E	N	E	N	E	N	E	N	E	N	E	N	E	N	E	N	E	N	Actual	
E	6	6	0	12	6	6	7	5	6	6	6	6	7	5	7	5	6	6	7	5	7	5	12	E
N	34	49	3	80	35	48	12	71	19	64	27	56	46	37	24	59	28	55	27	56	11	72	83	N
	40	55	3	92	41	54	19	76	25	70	33	62	53	42	31	64	34	61	34	61	18	77	95	

Source: the author

6. Conclusion and future studies

This paper aimed at presenting the viability in applying social economic attributes to classification algorithms regarding students at risk of dropping out from IFES, Guarapari campus, undergrad Business Administration course. This paper evaluated the algorithms from a open Data Mining software through three types of experiments that are available in the tool, for two different databases, with and without school attendance information, where the last one showed the worse, but still satisfactory, performance.

These computational techniques can aid school administrators in planning pedagogical and administrative actions in direct and assertive manners. Besides, by combining social economic and school performance data, the algorithms tend to show even better results, allowing for a continued practice of students' dropout prevention. In future studies, we expected to apply these experiments in other undergrad courses on the Institute, expanding the observations and research into these technological possibilities. Therefore, with a larger amount of data available, we also expect to identify other potential uses of Data Mining in academic administration.

7. References

[1] A. Nora, E. Barlow, G. Crisp, "Student persistence and degree attainment beyond the first year in college", College student retention: Formula for success, 2005, pp. 129-153.

[2] A.F. Cabrera, A. Nora, M.B. Castañeda, "The role of finances in the persistence process: a structural model", Research in Higher Education, v. 33, n. 5, 1992.

[3] A.W. Astin, "College student retention: Formula for student success", Rowman & Littlefield Publishers, 2012.

[4] A.W. Astin, "Student Involvement: A developmental theory for higher education", Journal of College Student Personnel, 1984.

[5] B. Kipnis, "A pesquisa institucional e a educação superior brasileira: um estudo de caso longitudinal da evasão", Linhas Críticas, v. 6, n. 11, 2000, pp. 109-130.

[6] BRAZIL, "Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas", MEC, 1996.

[7] C.A. Biazus, "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", Doctoral Thesis in Product Engineering, Federal University of Santa Catarina, 2004.

[8] C.A.S. Baggi, D.A. Lopes, "A evasão e avaliação institucional no ensino superior: Uma discussão bibliográfica.", Avaliação: Revista da Avaliação da Educação Superior, v. 16, n. 2, 2011, pp. 355-374.

[9] E.T. Pascarella, “Student-faculty informal contact and college outcomes”, Review of Educational Research, v. 50, n.4, 1980.

[10] F. MacKinnon-Slaney, “The adult persistence in learning model: A road map to counseling services for adult learners”, Journal of Counseling & development, v. 72, n. 3, 1994, pp. 268-275.

[11] F.P. Schargel, J. Smink, “Estratégias para auxiliar o problema de evasão escolar”, Dunya, v. 282, 2002.

[12] J. Han, J. Pei, M. Kamber, “Data mining: concepts and techniques", Elsevier, 2011.

[13] J.M. Braxton, A.S. Hirstchi, S.A. McClendon, “Understanding and Reducing College Student Departure”, Ashe-ERIC Higher Education Report, v. 30, n. 3, 2004.

[14] J.P. Bean, “Dropout and turnover: The synthesis and test of a causal model of student attition”, Research in Higher Education, v. 12, 1980.

[15] J.P. Bean, B.S. Metzner, “A conceptual model of nontraditional undergraduate student attition", Review of Educational Research, v. 55, 1985.

[16] L.M.B. Manhães, S.M.S. Cruz, R.J.M. Costa, J. Zavaleta, G. Zimbrão, “Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa”, Brazilian Symposium on Information Systems, 2012.

[17] L.M.B. Manhães, S.M.S. Cruz, R.J.M. Costa, J. Zavaleta, G. Zimbrão, “Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados”, Brazilian Symposium on Computers in Education, 2012.

[18] N.P. Gaioso, “O fenômeno da evasão escolar na educação superior no Brasil”, Masters in Education Dissertation, Catholic University of Brasília, 2005.

[19] O.V. Nascimento, “Cem anos do ensino profissional no Brasil”, Ipbex, 2007.

[20] R. Baker, K. Yacef, “The State of Educational Data Mining in 2009: A Review and Future Visions”, Journal of Educational Data Mining, v. 1, n. 1, 2009, pp. 3-17.

[21] R. Baker, S. Isotani, A. Carvalho, “Mineração de Dados Educacionais: Oportunidades para o Brasil”, Brazilian Journal of Informatics in Education, v. 19, n. 2, 2009.

[22] R.L.L. Silva Filho, P.R. Motejanas, O. Hipólito, M.B. Lobo, “A evasão no Ensino Superior Brasileiro. Instituto Lobo para o Desenvolvimento da Educação, da Ciência e da Tecnologia.” Cadernos de Pesquisa,

v. 37, n. 132, 2007.

[23] R.S. Freitas, "A ocorrência da evasão do ensino superior - Uma análise das diferentes formas de mensurar", Masters in Education Dissertation, State University of Campinas, 2016.

[24] V. Tinto, "Classrooms as communities: exploring the educational character of student persistence", Journal of Higher Education, v. 68, n. 6, 1997.

[25] V. Tinto, "Dropout from higher education: a theoretical syntehesis of recente research", Review of Educational Research, v. 45, n. 1, 1975.

[26] V. Tinto, "Leaving college: Rethinking the causes and cures of students attrition", University of Chicago, n. 2, 1993.

[27] V. Tinto, "Research and practice of student retention: what next?", Journal of College Student Retention: Research, Theory and Practice, v. 8, n. 1, 2006.

[28] W.G. Spady, "Dropouts from Higher Education: An interdisciplinary review and synthesis", Interchange, v.1, 1970.

[29] W.S. Swail, "The art of student retention: A handbook for practitioners and administrators", Educational Policy Institute, Texas Higher Education Coordinating Board 20th Annual Recruitment and Retention Conference Austin, 2004, pp. 1-39.