Communication Papers of the Federated Conference on Computer Science and Information Systems pp. 3–10
DOI: 10.15439/2019F265
ISSN 2300-5963 ACSIS, Vol. 20

Comparative Analysis of Data Mining Algorithms Applied to the Context of School Dropout

Nathanael Oliveira Vasconcelos*, Methanias Colaço Júnior†, Thiago S. Almeida‡ and Victor Matheus da Silva §
*‡§ Federal University of Sergipe, Brazil
*†‡ Postgraduate Program in Computer Science - PROCC
*Email: nathanvasc@gmail.com
†Competitive Intelligence Research and Practice Group – NUPIC
†Email: mjrse@hotmail.com
‡ Email: thiago7a@gmail.com
§Email: victormsti@gmail.com

***Abstract*—Students' dropout is certainly one of the major problems that afflict educational institutions, the losses caused by the student's abandonment are social, academic and economic waste. The quest for its causes has been subject of work and educational research around the world. Several organizations seek strategic decisions to control the dropout rate. This work's goal is to evaluate the effectiveness of the most used data mining algorithms in the education area. An "in vivo" controlled experiment was planned and performed to compare the efficacy selected classifiers. The Random Forest and SVM algorithms have stood out in this context, having, statistically similar accuracy (80.36%, 81.18%), precision (80.79%, 80.25%), recall (76.50%, 77.51%) and f-measure (78.86%, 78.81%) averages. The results showed evidence of significant differences between the algorithms, and also showed that, although the SVM had the best metric of accuracy and recall, it results were statistically similar with Random Forest results.**

I. Introduction

According to the 2015 Higher Education Census in Brazil, 11% of the students entering the undergraduate program in 2010 dropped out in the first year. By 2014, almost half (49%) of students had left the courses they had opted for in 2010 [8]. Thus, dropout is certainly one of the major problems that afflict educational institutions in general. The search for its causes has been the object of many studies and educational researches [32] [22] [21][14].

Several brazilian governmental organizations such as the REUNI Program - Reestruturação e Expansão das Universidades Federais (Restructuring and Expansion of Federal Universities) - and the TAM - Termo de Acordo de Metas e Compromissos (Goals and Commitments Agreement Term) - seek for strategic decisions that try to control the dropout rate [32]. The losses caused by the student's abandonment are social, academic and economic waste. In the public sector, resources are invested without due return, in parallel, in the private sector, evasion rates represent a significant loss of revenue. Consequently, there is the need to increase the understanding of the problem and its causes, by adopting more effective measures to identify and understand the main factors that can contribute to student failure.

A very promising information gatthering alternative is the use of "knowledge discovery in databases" and the use of "data mining techniques in education", also called Educational Data Mining (EDM) [22].

EDM is defined by "The Educational Data Mining" community website[1] as an emerging discipline, concerned with the development of methods to explore the unique types of data that come from educational environment and use those methods to better understand the students and the characteristics of their learning.

Similarly, it is possible to mine data from students in order to identify relationships among the various factors that lead them to abandon the course. However, predicting school dropout is a multifactorial problem that includes several variables such as family, social, economic, and personal factors [27].

This work's goal is to evaluate the effectiveness of the most used data mining algorithms in the educational area for the prediction of students on the verge of academic dropout in the context of a public higher education institution. The results show evidence that the Random Forest and SVM algorithms have stood out in this context, having, statistically similar accuracy (80.36%, 81.18%), precision (80.79%, 80.25%), recall (76.50%, 77.51%) and f-measure (78.86%, 78.81%) averages.

This paper is organized as follows. Section II presents the related works. Section III presents the methodology used in this research. Section IV presents the required theoretical concepts to understand the research. Section V presents experiment definition and planning, followed by the presentation of its operation in Section VI. Section VII presents the results, and final considerations can be found in Section VIII.

II. Related Works

Recently, a reasonable number of researches have been conducted to apply data mining techniques in the education area, in order to classify and predict student performance in various education institutes [10] [22] [21] [17] [20]. The use of such techniques in education is promising due to the ammount of oportunities in this area [1].

[1]www.educationaldatamining.org

Iam-on and Boongoen [14] present as a study case the Mae Fah Luang University, in Thailand, by using EDM models and proposing a new data transformation approach to improve the accuracy of conventional classifiers aiming at the disseminating of interesting patterns with a higher accuracy. Their works contributed to predict students' performance and possible dropouts, based on their pre-university characteristics, admission details, and initial academic performance at the university. The limitation to their model is the complexity and the required time, so it may not work well with larger datasets.

Dekker et al. [10] were able to identify, in the first school year, the students who presented the highest risk of dropout. The study considered several students' data and obtained accuracy between 75% and 80% using a tree decision classifier.

Márquez-Vera et al. [22] propose the application of data mining techniques to predict school failure and dropout in a case study with data from 670 high school students in Zacatecas, Mexico. The acuracies obtained ranged from 75% to 98%, considering ten classifiers. In this study, students' scores were used with greater emphasis, in relation to other attributes. The authors conclude that classification algorithms can be used successfully in order to predict students' development. It is worth noting that despite having a high accuracy in some trials, the context applied is different from the one proposed in this work.

Some EDM studies in the scope of the Brazilian school dropout are highlighted, and we will discuss them in the next paragraphs.

Manhães et al. [21] compared 6 classifying algorithms and found problems with students who can not complete their undergraduate courses. The data sample is composed of 7304 students from higher education course at UFRJ - Universidade Federal do Rio de Janeiro. The data are classified into three classes: students who completed the course obtaining the diploma, students who could not complete the course, and students who had active enrollment after the average deadline for the conclusion of the undergraduate course at the federal institution. The study obtained an accuracy precision of around 80%.

Pascal et al. [17] addresses the dropout rate in a graduation course in a public higher education institution, considering only the Business Management and Zootechny courses. The research uses several machine-learning methods for prediction, and its tests have reached an accuracy higher than 70 % in students dropout prediction.

Machado et al. [20] presented in the article "Bibliometric study in data mining and school dropout", a bibliometric survey of articles published between 2005 and March 2015 that address data mining and school dropout issues. The survey was conducted by using the *Scopus*, *Web Of Science*, *Science Direct* and *Scielo* databases.

The search had as result 16 articles from *Scopus* database, 6 from *Web of Science*, 3 from *Sience Direct* and none from *Scielo* databases. Therefore, 24 scientific articles were part of this study scope.

Machado et al. [20] also listed nine data mining methods found in their research, among them *Decision Trees*, *Neural Network*, *Rule Induction* and *Support Vector Machine*.

In Table I, it is possible to observe a list with the previously cited related works, their classifiers and the obtained accuracies. It is important to note that the papers use different databases for their studies, making it impossible to compare the results directly. However, some of these studies have also used data from a public higher education institution, and along with our article, may help in a secondary general analysis.

It is worth mentioning that there were no studies that performed a comparative analysis of algorithms applied to the school dropout context considering an experimental approach with statistical validation of the significance of data, as proposed in this paper. A robust knowledge base can only be generated with the replications of real controlled experiments that statistically validates their work, which can serve as input for real data meta-analyzes.

III. METHODOLOGY

The methodology used in this work, in terms of classification, consists of an exploratory research [29], as a literature review was conducted with systematic approaches, in which the attributes used to generate the files used by the mining tool were defined. The selection of attributes was performed after analysis of the related works, like work [22], as well as by considering the attributes available in the database.

In addition, we used the attribute selection algorithm Random Forest, which evaluates the predictive value of each attribute individually, generating a ranking in which those attributes that have more relation with the class and less correlation with the other attributes receive higher scores [31].

After defining the attributes, mining models were generated, in order to make a comparison between the used algorithms. The main classifiers found in the EDM works [25][22] are: Decision Tree [5], Naive Bayes [28], Nearest Neighbor (KNN) [6], Neural Networks (MLP) [24], Support Vector Machines (SVM) [7] and Ensemble Methods (Random Forest)[4]. According to Xindong et al. [34] these algorithms are among the most used ones in data mining.

Finally, to achieve this research main goal and subsequent data collection, a controlled "in vivo" experiment was proposed and carried out, which involved the database of a public higher education institution. According to Wohlin et al.[33], experimentation is not a simple task, as it involves preparing, conducting and analyzing data correctly. The authors highlight the control of subjects, objects and instrumentation as one of the main advantages of the experimentation, which makes it possible to draw more general conclusions on the investigated subject.

Other advantages include the ability to perform statistical analyzes by using hypothesis testing methods and opportunities for replication. Juristo et al. [16] also state that scientific research can not be based on commercial opinions or interests. Scientific investigations are represented by studies based on observation and/or experimentation with the real world and

TABLE I
LIST OF THE MAIN RELATED WORKS, CLASSIFIERS USED AND ACCURACIES

Author	Classifiers	Accuracies
[10]	OneR, CART, Decision Tree, Naive Bayes, Net Logit, JRip, Random Forest	75% to 80%
[22]	JRip, NNge, OneR, Prism, Ridor, Decision Tree, SimpleCart, ADTree, RandomTree, REPTree	75% to 99%
[21]	Decision Tree, Support Vector machine (SVM), AdaBoost, Naïve Bayes, SimpleCart, MLP	Average = 80%
[17]	Decision Tree, KNN, CART, Naive Bayes, MLP	Average = 74%

their measurable behaviors, as in this research. These aspects should also be taken into account in the construction and evaluation of algorithms and software.

In the execution of the experiment, with the definition of the algorithms and attributes, the Python language and its libraries were used, with this, knowledge models were generated in order to perform algorithm tests and compare the effectiveness. In this context, this work was also classified as laboratory and experimental, due to the planning and execution of a controlled experiment.

To assist the calculations and to verify if there were significant differences in the algorithms efficiency, the Statistical Package for Social Science - SPSS [15] tool was used for data analysis, applying basic and advanced statistical techniques. The SPSS is a statistical software used internationally for many decades, since its versions for large computers [23].

In summary, the experiment can be divided into four main stages: planning with a selection of attributes; data cleaning operation, dataset generation and data collection; comparison of algorithms; and finally the results analysis. The experiment in question is detailed in Sections V and VI.

IV. CONCEPTUAL BASE

In this section, some concepts that are necessary for the understanding of this work are presented.

A. Data Classification

Classification is the process of associating specific objects (instances) into a set of categories (classes or concepts), based on their object properties. Classification is a procedure where individual items are placed in groups based on information derived from characteristics inherent in the items and based on a training set previously labeled [11]. The algorithms used in this research are cited in Section III.

B. Matrix of Confusion

Among the various ways of evaluating a classifier's predictive ability to determine the class of multiple records, the confusion matrix is the simplest of these forms [12].

For n classes, the confusion matrix is a dimension table $n \times n$. For each possible classification, there is a corresponding row and column, it means, the values of the classifications will be distributed in the matrix according to the results, thus generating the confusion matrix for the prepared classifications [3]. The rows correspond to the correct classifications, and the columns represent the classifications performed by the classifier [13] .

When there are only two classes, one is considered *positive* (in the context of this work, "Evaded") and the other as *negative* [13]. Thus, we can have four possible outcomes:

- *True Positive* (TP): a positive class instance is correctly classified as positive;
- *Negative* (FN): a positive class instance is incorrectly classified as negative;
- *True Negative* (TN): a negative class instance can be correctly denoted as negative;
- *Positive* (FP): a negative class instance is incorrectly classified as positive.

C. Quality Metrics

In this work, the accuracy, recall, precision and F-measure metrics were used.

1) Accuracy: It is the percentage of instances sorted correctly.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

2) Recall: It is the percentage of instances that were correctly classified as positive.

$$recall = \frac{TP}{TP + FN} \tag{2}$$

3) Precision: It is the percentage of instances rated positive that are really positive.

$$precision = \frac{TP}{TP + FP} \tag{3}$$

4) F-Measure : Also known as harmonic measure, because it combines precision and recall, evenly weighting.

$$F - measure = \frac{2.precision.recall}{precise + recall} \tag{4}$$

V. EXPERIMENT PLANNING

In this and the next two sections, this work is presented as an experimental process. It follows the [33] guidelines. This section will focus on goal setting and experiment planning. Figure 1 illustrates the steps of the work, this section will focus on step 1, that is, goal setting and experiment planning.

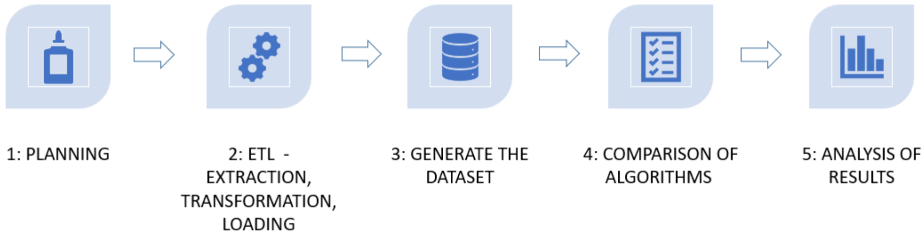

Fig. 1. Stages of the work.

A. Goal Setting

The goal of this study is to evaluate the main classifiers found in EDM, identifying the best algorithm in terms of effectiveness, focusing on school dropout, within a higher education institution. The major models will be used to form a predictive metamodel that will help with dropout management.

The experiment will target undergraduate students from a higher education institution. The goal was formalized by using the GQM model proposed by [3]: **Analyze**, through a controlled experiment, the main algorithms of data mining applied to the context of education, with **focus** on school dropout, **for the purpose** of confirming and/or identifying the best algorithm in terms of efficacy, **with respect to** accuracy, recall, precision and f-measure, **from the point of view** of researchers and data analytics professionals, **in the context** of data on the dropout rate of a public institution of higher education.

1) Context Selection: The experiment was "in vivo" and considered the data of students of all undergraduate courses, from a public higher education institution, with admittance between 2003, year in which the first undergraduate courses began, and 2017. The year/period of the institution during the period of preparation of this article was 2018/1, which, for this reason, was not part of this study. Data selection took into account personal, academic, and social attributes.

2) Hypothesis Formulation: To guide the study, the following research question was elaborated, whose answer aims to fulfill the objective of the work. In the context of the School Dropout Rates in a higher education institution, among the algorithms selected in the EDM area, which is the best in terms of Efficacy?

To evaluate this question, we used four metrics: Accuracy (1), Recall (2), Precision (3) and F-Measure (4). Thus, with the objectives and metrics defined, the following hypothesis will be considered (for each metric):

- H_0: The algorithms$_{(1,2..n)}$ have the same averages for the metric.
  μ1(metric) = μ2 (metric) ... = μ n (metric);
- H_1: The algorithms$_{(1,2..n)}$ have different averages for the metric.
  μ1 (metric) $\neq$ μ2 (metric) ... $\neq$ μ n (metric);

3) Selecting Participants: All undergraduate students of a higher education institution in all the complete academic periods were considered. This selected institution is public and has several courses of different levels. The database analyzed was the SIGAA - Sistema Integrado de Gestão de Atividades Acadêmicas (Integrated Academic Activities Management System), which stores the entire academic life of the institution's students. The institution provided the database for the experiment in question.

4) Independent variables: For the classification task, we considered 17 attributes, from the base described in subsection V-A3, which are presented in the Table II. The used algorithms are: Decision Tree, Naive Bayes, k-Nearest Neighbor (KNN), Neural Networks (MLP), Support Vector Machine (SVM) and Ensemble Methods (Random Forest), with the parameters presented in the Table III.

TABLE II
ATTRIBUTES CONSIDERED FOR ANALYSIS

Attribute	Description
sexo	Student gender
idade	Student age at the beginning of the course
inst_seg_grau	High school institution type
raca	Student Ethnicity
est_civil	Marital status
qtd_tranc	Number of stopouts in the course
reab_matricula	Indicates whether the student has re-enrolled in the course
qtd_ap_med_p	Average number of courses approved per period
qtd_ap_1p	Number of courses approved in the first period
qtd_rep_med_p	Average number of failed subjects per period
qtd_rep_1p	Number of failed subjects in the first period
qtd_per_cur	Number of periods attended by the student
cra	Academic performance coefficient
perc_aprov	Percentage of subjects approved by the student
media_geral	Overall grade of the student in the course
media_Faltas	Average student absences in the course
cotista	Indicates whether the student entered the course by quota system

5) Dependent Variables: Accuracy (1), Recall (2), Precision (3) and F-Measure (4).

6) Experiment Design: After the preprocessing, which consisted in the removal of records that were very different from the average, 6672 instances were selected, which represent all undergraduate students of the institution, within the period

TABLE III
USED PARAMETERS BY ALGORITHM.

Algorithm	KNN	Random Forest	Naive Bayes	SVM	Decision Tree	MLP
n_neighbors	25	-	-	-	-	-
random_state	-	0	-	-	0	-
criterion	-	entropy -	-	-	gini	-
n_estimators	-	75 -	-	-	-	-
max_depth	-	10 -	-	-	None	-
n_jobs	-	-1 -	-	-	-	-
max_features	-	0.3 -	-	-	None	-
bootstrap	-	-	-	-	-	-
C	-	-	-	0.001	-	-
cache_size	-	-	-	200	-	-
class_weight	-	-	-	None	None	-
max_iter	-	-	-	-1	-	200
probability	-	-	-	False	-	-
random_state	-	-	-	None	-	1
shrinking	-	-	-	True	-	-
tol	-	-	-	0.001	-	0.0001
verbose	-	-	-	False	-	False
coef0	-	-	-	0.0	-	-
decision_function_shape	-	-	-	ovr	-	-
degree	-	-	-	3	-	-
gamma	-	-	-	1	-	-
kernel	-	-	-	poly	-	-
max_leaf_nodes	-	-	-	-	None	-
min_impurity_decrease	-	-	-	-	0.0	-
min_impurity_split	-	-	-	-	None	-
min_samples_leaf	-	-	-	-	1	-
min_samples_split	-	-	-	-	39	-
min_weight_fraction_leaf	-	-	-	-	0.0	-
presort	-	-	-	-	False	-
splitter	-	-	-	-	best	-
activation	-	-	-	-	-	logistic
alpha	-	-	1	-	-	1e-05
batch_size	-	-	-	-	-	auto
beta_1	-	-	-	-	-	0.9
beta_2	-	-	-	-	-	0.999
early_stopping	-	-	-	-	-	False
epsilon	-	-	-	-	-	1e-08
hidden_layer_sizes	-	-	-	-	-	(3, 2)
learning_rate	-	-	-	-	-	constant
learning_rate_init	-	-	-	-	-	0.002
momentum	-	-	-	-	-	0.9
shuffle	-	-	-	-	-	True
solver	-	-	-	-	-	adam

previously mentioned. Of the selected data, 3212 (48.1%) represents dropped out students and 3460 (51.9%) represent active students.

One of the metrics used in this work was the accuracy, which requires the balancing of class data. Since our base is already balanced [19], it was not necessary to plan the adoption of a balancing method.

The 10-fold Cross-validation approach was used, where the data are divided into 10 parts, maintaining their proportions. Thus, 10 tests are performed, in which part of the data is separated to be tested later and the others are used to be trained.

7) Instrumentation: For the data mining process, the Python language and its libraries was used, which has several machine learning algorithms that can be used to extract relevant information from a database. According to the 16th analysis software usage and data mining annuary [26], Python was considered the most used programming language by Data Mining and Big Data professional community.

Python has many reasons for attracting interest as a language for data analysis: it is open-source and free of cost, it has a varied set of libraries to work with several areas allowing performance comparison between the algorithms and presenting several resources for data analysis. In addition it offers a simple and objective syntax that allows the programmer to focus on the problem to be solved without worrying so much about details of implementations.

The data used for the analysis comes from SIGAA, which has the PostgreSQL as SGBD. An ETL (Extract, Transform and Load) was created to extrac, clean and load the data in a specific Data Warehouse, which is the basis for generation of knowledge models, taking into account the variables detailed in subsection V-A4.

VI. Experiment Operation

A. Preparation

It consisted of implementing the ETL implemented to load the Data Warehouse. The data were submitted to pre-processing, in which records with different values of the average (outlies) were removed. In this step, it was also done the transformation of some attributes, in which the "One Hot" approach was applied, consisting of representing a categorical variable of binary form. This process is represented in steps 2 and 3 of Figure 1.

B. Execution

It consisted in performing the classifying process in the data of the students, planned in section V-A6, for each selected mining algorithm, by using the dictionary discussed in subsection V-A4. Step 4 of Figure 1.

C. Data Validation

Four types of statistical tests, Shapiro-Wilk Test, Levene Test, Anova Test and Tukey Test were used as an aid to analysis, interpretation and validation - step 5 of Figure 1.

The Anova Test was used because it was necessary to compare more than two groups of values. Since this test has the assumptions that the distribution must be Normal and that there is homoscedasticity between the treatments (homogeneous variances) [9], the Shapiro-Wilk [30] Test was used for the Normality test and Levene's test [18] for the homoscedasticity test.

The Anova Test shows that at least one algorithm differs from the others, but it is not possible to say which one is more dissimilar. For this, the Tukey test was used, which according to Angels [2], allows to test any contrast, always, between two averages of treatments, being possible to verify which are statistically the same or different.

All statistical tests were done using the SPSS Tool - IBM [15].

VII. Results

After the execution of the algorithms using the 10-cross-validation approach, the results of the classifications were obtained. In Table IV and in Figure 2, the averages obtained by each algorithm with all the attributes are presented.

TABLE IV
Comparative of the metrics of the algorithms with all attributes.

Algorithms	Accuracy	Precision	Recall	F Measure
Knn	77,52%	77,56	77,51%	77,46%
Random Forest	80,63%	80,28%	80,33%	80,3%
Naive Bayes	77,13%	76,84%	77,45%	76,9%
SVM	81,01%	77,88%	78,02%	77,91%
Decision tree	77,62%	77,44%	77,7%	77,49%
MLP	79,15%	79,3%	79,25%	79,25%

By using the Random Forest algorithm for attribute selection, it is possible to notice that some have more relevance and others could be eliminated without influencing the results. The

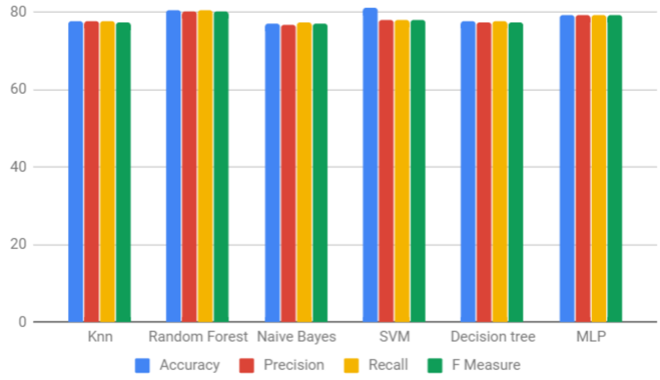

Fig. 2. Comparative chart of the metrics of the algorithms with all attributes.

4 attributes of greatest relevance had the following order, starting from the most relevant: "idade", "sexo", "qtd_ap1p" and "media_geral". The attributes that presented lower relevance were "qtd_rep_med_p" (average number of failed subjects per period), "qtd_per_cur" (percentage of finished courses) and "qtd_ap_med_p" (average number of courses approved perperiod).

This shows, for example, that the students' age and gender has influence on the problem and could be related with the responsabilities each of them have besides the studies. It also brings an alert for further analysis of the social areas of the institution, considering the gender and age group in the basic education that are being most affected. In addition, it is evident that the student's performance in the first period is quite relevant.

After selecting the attributes less relevant, the algorithms were executed again. The averages of each algorithm are presented in Table V and in Figure 3 below.

TABLE V
Comparative of the metrics of the algorithms with the selected attributes.

Algorithms	Accuracy	Precision	Recall	F-Measure
Knn	78,01	79,09	70,83	74,69
Random Forest	80,36	80,79	76,50	78,86
Naive Bayes	76,8	79,00	67,59	72,81
SVM	81,18	80,25	77,51	78,81
Decision tree	78,06	76,51	75,15	75,79
MLP	79,43	77,34	76,92	77,06

These results were used to respond the research question. The algorithms obtained distinct average accuracies and the SVM algorithm obtained the highest ones, followed by Random Forest, which achieved very simmilar averages and close to SVM's. However, it is not possible to make such assumptions without conclusive statistic evidence. For that reason, the Anova Test was applied to validate the hypotheses. At first, for having the assumptions of normality and homoscedasticity, the Shapiro-Wilk Test was carried out, followed by the Levene Test.

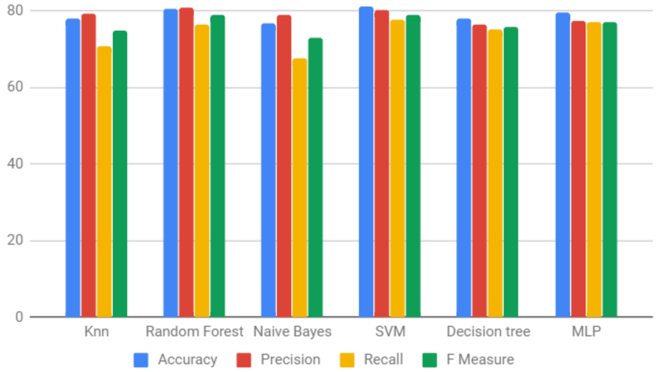

Fig. 3. Comparative chart of the metrics of the algorithms with the selected attributes.

A 0.05 significance level to the experiment was defined . When applying the Shapiro-Wilk Test for the normality analysis of the distribution of the data the p-values from Table VI were obtained. In which values above the level of significance were observed, leading to the conclusion that the distributions are normal.

TABLE VI
THE SHAPIRO-WILK TEST RESULTS FOR DATA NORMALITY ANALYSIS.

Algorithms	Accuracy	Precision	Recall	F-Measure
Knn	0,167	0,736	0,600	0,338
Random Forest	0,706	0,633	0,018	0,837
Naive Bayes	0,089	0,807	0,566	0,051
SVM	0,637	0,560	0,306	0,579
Decision tree	0,735	0,980	0,482	0,698
MLP	0,592	0,870	0,453	0,289

Then, the Levene Test was performed, obtaining the results presented in Table VII, that is, higher than the significance level adopted, validating the variances homogeneity assumption.

Once the assumptions were met, it was possible to apply the Anova test, in which p-values were significantly lower than the level of significance adopted, as can be seen in the table VII. In this way, the evidence of difference between the averages was confirmed, that is, the hypothesis (H_0) that the algorithms have the same accuracy was rejected within the context of the experiment.

TABLE VII
P-VALUES OF THE LEVENE AND ANOVA TESTS.

Metrics	Levene	Anova
Accuracy	0,807	$4,10^{-6}$
Precision	0,795	$3,02^{-6}$
Recall	0,978	$4,29^{-13}$
F-Measure	0,469	$4,01^{-9}$

The Anova Test evidence that at least one algorithm differs from the others, but it is not possible to affirm wich one. For this, the Tukey Test was used, because it allows to test any contrast between two averages treatment, making it possible to verify which are statistically different or equal, accordingly to [2].

Figure 4, following, presents the average accuracies grouped algorithms, forming four homogeneous groups. By analyzing the figure we, can see that th highest average belongs to the SVM algorithm (81.18%). However, considering the significance level of 5% it is possible to say that SVM and Random Forest obtained the same average accuracy. Naive Bayes achieved the lowest average (76.50%).

Algorithm	Subset for alpha = 0.05			
	1	2	3	4
Naive Bayes	76,80			
KNN		78,01		
Decision Tree		78,06	78,06	
MLP			79,43	
Random Forest				80,36
SVM				81,18
Sig.	1	0,865	0,101	0,597

Fig. 4. Values obtained by the Tukey's Test on Accuracy.

Similarly, the Figure 5 shows the result of the Tukey test of the F-Measure. To avoid repetition, the Tukey's test of precision and recall will be omitted in this work, since the F-Measure makes the combination of these metrics. The results show that the Random Forest, SVM and MLP algorithms presented similar averages for the F-Measure, being 78.86%, 78.81% and 75.99%, respectively. Naive Bayes had the lowest average with 72.81%.

Algorithm	Subset for alpha = 0.05			
	1	2	3	4
Naive Bayes	72,81			
KNN	74,69	74,69		
Decision Tree	75,79	75,79	75,79	
MLP		75,99	75,99	75,99
SVM			78,81	78,81
Random Forest				78,86
Sig.	0,056	0,802	0,051	0,073

Fig. 5. Values obtained by the Tukey Test's on F-Measure.

A. *Threats to Validity*

Although the results of the experiment were satisfactory, it presents threats to its validity that should be commented on.

Threats to internal validity: The current academic system has been present in the Institution since 2017, which inherited the basis of the academic system legacy, with several inconsistent information, mainly until the middle of 2007. This threat was mitigated with the accomplishment of the cleaning process that decreased the likelihood of using incorrect older information.

Threats to construction validity: In the experiment of this article, the institution did not possess any very relevant information regarding dropout, like, for example, socioeconomic data of the students and their relatives. The inclusion of this information can influence the performance of the algorithms, increasing their efficiency. To mitigate this threat and increase the Decision Support System yet to be developed, we will suggest that this information be gathered by the institution and taken into account in future work.

VIII. Final Considerations

The main contribution of this work was to evaluate six major algorithms most commonly used in the context of EDM in terms of Accuracy, Precision, Recall and F-Measure for identify identifying the factors that influence school dropout.

The work was consolidated with the conduction of a controlled experiment in which the results showed that there are significant differences between the algorithms used, and that the algorithms Random Forest and SVM have stood out in this context, possessing, statistically, similar Accuracy (80.36%, 81.18%), Precision (80.79%, 80.25%), Recall (76.50%, 77.51%) and F-measure (78.86%, 78.81%) averages. The results shows evidence of significant differences between the algorithms, and that although the Random Forest and the SVM has the best metrics evaluated, its results are statistically similar with MLP results.

Besides that, the published works on the subject have some scientific gaps if we consider that there was no rigorous validation of the results, allowing a more assertive combination of experimental evidence. In this context, this paper validates its results and confirms some previous evidence found in the works described in [22] e [10].

Finally, as future work, we intend to analyze the algorithms in other levels of education, as well as add other types of variables for analysis, such as socioeconomic information. In addition, we intend to develop a predictive system for teaching management, which can help in decision making process to combat school dropout and school retention.

References

[1] A. M. Ahmed, A. Rizaner, and A. H. Ulusoy. Using data mining to predict instructor performance. *Procedia Computer Science*, 102:137–142, 2016.

[2] A. Anjos. Análise de variância. *Universidade Federal do Paraná, Departamento de Estatística - UFPR, Curitiba*, page Capítulo 7, 2009.

[3] V. R. Basili and D. M. Weiss. A methodology for collecting valid software engineering data. Technical report, NAVAL RESEARCH LAB WASHINGTON DC, 1983.

[4] L. Breiman. Machine learning. *Kluwer Academic Publishers*, pages 5–32, 2001.

[5] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classication and regression trees*. Monterey, CA: Wadsworth and Brooks, 1984.

[6] G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, 30(4):891–927, Jul 2016.

[7] C. Cortes and V. Vapnik. Support-vector networks. machine learning, 20. pages 273–297, 1995.

[8] I. N. de Estudos e Pesquisas Educacionais Anísio Teixeira – Inep. Censo da educação superior 2015. 2015.

[9] A. Field. *Descobrindo a estatística usando o SPSS*. 2.ed. Porto Alegre: Artned, 2009.

[10] D. G., P. M., and V. J. Predicting students drop out: A case study. *Proceedings of the International Conference on Educational Data Mining*, pages 41–50, 2009.

[11] F. Gorunescu. *Data Mining: Concepts, models and techniques*, volume 12. Springer Science & Business Media, 2011.

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

[13] D. J. Hand, H. Mannila, and P. Smyth. *Principles of data mining*. MIT press, 2001.

[14] N. Iam-On and T. Boongoen. Improved student dropout prediction in thai university using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics*, 8(2):497–510, 2017.

[15] IBM. Spss. *IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp*, 2017.

[16] N. Juristo and A. Moreno. Software engineering experimentation. 2001.

[17] G. Kantorski, E. G. Flores, J. Schmitt, I. Hoffmann, and F. Barbosa. Predição da evasão em cursos de graduação em instituições públicas. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 27, page 906, 2016.

[18] H. Levene. Robust tests for equality of variances. *International Journal of Machine Learning and Cybernetics*, pages 278–292, 1960.

[19] E. Machado and L. Marcelo. Um estudo de limpeza em base de dados desbalanceada e com sobreposição de classes. *XXVII Congresso da Sociedade Brasileira de Computação*, 2007.

[20] R. D. Machado, E. O. B. Nara, J. N. C. Schreiber, and G. A. Schwingel. Estudo bibliométrico em mineração de dados e evasão escolar. *XI Congresso Nacional de Excelência em Gestão*, 2015.

[21] L. M. B. Manhães, S. Cruz, R. J. M. Costa, J. Zavaleta, and G. Zimbrão. Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. *Anais do VIII Simpósio Brasileiro de Sistemas de Informação, São Paulo*, 2012.

[22] C. Márquez-Vera, C. R. Morales, and S. V. Soto. Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 8(1):7–14, 2013.

[23] E. Mundstock. Introdução à análise estatística utilizando o spss 13.0. cadernos de matemática e estatística série b. 2006.

[24] A. Nürnberger, W. Pedrycz, and R. Kruse. *Handbook of data mining and knowledge discovery. Chapter data mining tasks and Methods: Classication: Neural network approaches*. New York, NY, USA: Oxford University Press, 2002.

[25] J. G. d. Oliveira Júnior et al. Identificação de padrões para a análise da evasão em cursos de graduação usando mineração de dados educacionais. Master's thesis, Universidade Tecnológica Federal do Paraná, 2015.

[26] K. S. Poll. Analytics, data mining software used. *https://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html*, 2015.

[27] A. Pradeep, S. Das, and J. J. Kizhekkethottam. Students dropout factor prediction using edm techniques. In *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*, pages 1–7. IEEE, 2015.

[28] S. J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[29] A. J. Severino. *Metodologia do trabalho científico*. Cortez editora, 2017.

[30] S. Shapiro and M. Wilk. An analysis of variance test for normality (complete samples). *International Journal of Machine Learning and Cybernetics*, 52:591–611, 1965.

[31] D. F. Silva and G. E. de Almeida Prado Alves Batista. Uma comparação experimental de métodos de imputação de valores desconhecidos. *ICMC - Instituto de Ciências Matemáticas e de Computação, São Paulo*, 2009.

[32] R. L. L. Silva Filho, P. R. Motejunas, O. Hipólito, and M. Lobo. A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, 37(132):641–659, 2007.

[33] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[34] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.