Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 15/Fev/2022; Camera ready: 31/Aug/2022; 1st round notif.: 11/Apr/2022; Edition review: 8/Sep/2022; New version: 24/May/2022; Available online: 27/Sep/2022; 2nd round notif.: 13/Aug/2022; Published: 27/Sep/2022;

Educational data mining to support identification and prevention of academic retention and dropout: a case study in introductory programming

Murillo Guimarães Carneiro
Faculty of Computing
Federal University of Uberlândia
ORCID:0000-0002-2915-8990
mgcarneiro@ufu.br

Bruna Luiza Dutra
Faculty of Computing
Federal University of Uberlândia
ORCID:0000-0002-8614-8019
brunaluiza033@hotmail.com

José Gustavo S. Paiva
Faculty of Computing
Federal University of Uberlândia
ORCID:0000-0003-3228-6974
gustavo@ufu.br

Paulo Henrique Ribeiro Gabriel
Faculty of Computing
Federal University of Uberlândia
ORCID:0000-0002-5923-4181
phrg@ufu.br

Rafael Dias Araújo
Faculty of Computing
Federal University of Uberlândia
ORCID:/0000-0003-0545-2519
rafael.araujo@ufu.br

Abstract

Several works in the literature emphasized data mining as efficient tools to identify factors related to retention and dropout in higher education. However, most of these works do not discuss if (or how) such factors may effectively contribute to decrease such rates. This article presents a data mining approach conceived to identify students at retention risk in a course of Intro to Computer Programming as well as guide preventive interventions to help such students to overcome this situation. Our results indicated an averaged predictive performance superior to 80% in both accuracy and F1 when identifying factors related to the retention. Moreover, during the two years of the project execution, the annual success rates in the course were the highest in comparison to the last five years.

Keywords: *Educational data mining; Higher education; Retention prevention; Dropout prevention; Academic analytics; Machine learning; Data classification.*

Cite as: Carneiro, M. G., Dutra, B. L., Paiva, J. G. S., Gabriel, P. H. R., Araújo, R. D. (2022). Educational data mining to support identification and prevention of academic retention and dropout: a case study in introductory programming. Revista Brasileira de Informática na Educação, 30, 379–395. DOI: 10.5753/rbie.2022.2518.

1 Introduction

The large number of variables involved in educational decision-making processes has motivated the application of data mining techniques in this context, which is known as Educational Data Mining (EDM). Indeed, various EDM models act as catalysts for an adequate understanding and response to several educational problems due to their descriptive and predictive abilities. In the specific case of programming courses in higher education, data mining has played an essential role in understanding the phenomenon of high failure and dropout rates, which represents a major challenge in introductory programming courses, commonly known as ICP (Introduction to Computer Programming or Computer Science 1), and other computing courses in general (Petersen, Craig, Campbell, & Tafliovich, 2016; Silva, Borges, Ferreira, Santos, & Andrade, 2021; Neves et al., 2021).

Predictive modeling, the object of study of this paper, has already drawn attention since the first publications focused on distance education (Kotsiantis, Pierrakeas, & Pintelas, 2003). In Brazil, the topic took longer to become popular (Baker, Isotani, & Carvalho, 2011), with the first publications focused on both traditional learning (Manhães, Da Cruz, Costa, Zavaleta, & Zimbrão, 2012) and distance education (Gottardo, Kaestner, & Noronha, 2012). In the context of ICP, noteworthy efforts can be found in the literature on searching for factors that may or may not influence students' decision to continue in that course or that explain their failure (Horton & Craig, 2015; Petersen et al., 2016; Pappas, Giannakos, & Jaccheri, 2016; K. J. O. Santos, Menezes, de Carvalho, & Montesco, 2019; Hawlitschek, Köppen, Dietrich, & Zug, 2019; Oliveira, Ambrósio, Silva, Brancher, & Franco, 2020).

Nowadays, there is a growth in the number of works published both in Brazil and internationally on EDM in the context of school dropout, as shown in the literature review carried out by V. Santos, Saraiva, and Oliveira (2021). Most of them present evidence that data mining algorithms have good predictive power, allowing a better understanding of the problem. However, few studies evaluate the effect of this understanding in combating the phenomena — usually, they present rules generated by data mining algorithms and claim that they can help managers. Nevertheless, without evidence, these conclusions are nothing more than hypotheses; consequently, all the work related to data preparation and algorithm evaluation seems incomplete and insufficient to solve the central problem that motivated this entire process: the reduction in failure or dropout rates.

This paper presents a study conducted in a Bachelor's Degree in Information Systems course at the Federal University of Uberlândia, a Brazilian public institution, between 2018 and 2019. Such courses present high rates of evasion (Damasceno & Carneiro, 2018). Specifically, the methodology and results presented here refer to the ICP course, offered in the first semester of the course and with an average success rate of only 22% in the period of 2016-2017. Thus, the key questions that guided the investigation were:

- Research Question 1 (RQ1). Is it possible to develop a predictive model capable of revealing the factors that most contribute to the identification of students with a greater chance of failing the course?
- Research Question 2 (RQ2). Can such factors help guide practical actions to combat failure in the course?

The most significant contributions of this work consist of answering both questions. Initially, we trained a set of data mining algorithms to learn complex correlations among several information collected from the students, in order to identify the ones with high chance to fail. We evaluate and interpret the assumptions built by the learned models which may result in knowledge at the end of this process. Afterward, we developed a series of interventions guided for the at-risk students who were pointed out by our predictive model to combat the phenomenon of failure in the course. In summary, the annual success rates in this course reached the highest values since the creation of the degree in our campus.

We organize the rest of the paper as follows: Section 2 discusses the main related works; Section 3 presents the methodological details followed to conduct this investigation; Sections 4 and 5 bring the results and discussions about the predictive modeling and actions to prevent evasion and retention in our context; finally, Section 6 presents the conclusions of this work.

2 Related Work

This section presents the main related studies found in the literature, with a special focus on predictive modeling of failure or dropout in higher education. An extensive literature review was carried out covering dozens of works. There is a large number of researches focused on Virtual Learning Environments (VLE), such as Moodle (Gottardo et al., 2012), justified by the little contact between teachers and students in distance learning courses, and by the huge variety of attributes they provide. It is also worth mentioning some approaches that aim at predicting the undergraduate dropout rate (Manhães et al., 2012), although most studies are related to the detection of failure factors in one or a few specific courses (R. Santos, Pitangui, Vivas, & Assis, 2016).

The application of data mining in the educational context is widely explored in the EDM literature (Romero & Ventura, 2020), especially to predict students performance (Salloum, Alshurideh, Elnagar, & Shaalan, 2020) and to understand the factors that most impact in students success/failure. Most of techniques employ data regarding students economic and demographic information (Yu, DiGangi, Jannasch-Pennell, & Kaprolet, 2010), as well as examination results and course enrolment (Li, Ding, & Liu, 2020; Adekitan & Salau, 2019).

Several data mining techniques have been explored, including Decision Trees, Neural Networks and Random Forest (R. Santos et al., 2016; Palacios, Reyes-Suárez, Bearzotti, Leiva, & Marchant, 2021). Feature selection approaches are also employed to comprehend how each student information is related to his/her performance. (Carrano, de Albergaria, Infante, & Rocha, 2019). Finally, additional data mining tasks, such as clustering (Beltran, Xavier-Júnior, Barreto, & Oliveira Neto, 2019) and sequential pattern mining (Pimentel, Passos, Fernandes, & Goldschmidt, 2019) were also adopted with the aim of providing pre or post processing for the predictive models.

The most representative approaches related to this study are discussed below. Concerning the dropout identification, the works of Manhães et al. (2012) and Carrano et al. (2019) present, respectively, studies carried out with data from students from the Federal University of Rio de Janeiro and the Federal University of São João del-Rei, both Brazilian higher education institutions. While Manhães et al. (2012) developed their investigation on a database of students of the Civil Engineering course, Carrano et al. (2019) proposed to identify dropout considering all students enrolled at the university during a pre-defined period.

Other works also proposed the development of systems or platforms for monitoring and combating evasion (Beltran et al., 2019; Noetzold & de L. Pertile, 2021). As a result, these works demonstrate the promising ability of data mining techniques for evasion detection, but none of them presents results or discussions about confronting the phenomenon. In the context of the Information Systems course, Noetzold and de L. Pertile (2021) built a predictive model of school dropout based on decision trees and have shown that academic performance is a strong indicator of dropout and other factors such as the distance between student housing and the institution, or the improvement in the institution's infrastructure were not relevant for the dropout in that study.

Regarding the predictive modeling of failure, some works deserve to be highlighted (Kampff, Ferreira, Reategui, & Lima, 2014; Cambruzzi, Rigo, & Barbosa, 2015; Jayaprakash, Moody, Lauría, Regan, & Baron, 2014), which, unlike the vast majority of studies classified in this group, present their results when the failure factors identified are used to combat the phenomenon. In the works of Kampff et al. (2014) and Cambruzzi et al. (2015), which include distance education through a VLE, the interventions against failure took place by periodically sending alerts (messages) to teachers and students. In the work of Jayaprakash et al. (2014), the only one that considered the context of regular education, the intervention also took place through alerts. In these three works, it was demonstrated that the alerts derived from data mining systems can help to reduce the number of students who fail in the courses.

Like the work of Jayaprakash et al. (2014), the investigation presented in this article is also designed towards a face-to-face course. However, the target course of this study has a failure rate of 78%, which is much higher than the 7% dropout estimate that were considered in the place where that study was conducted. In this way, previous experience with past classes of the same course tells us that just alerting students is not enough. By the contrary, more than alerts, data mining will need to guide efforts towards a series of pedagogical and support interventions for students identified at risk. In this sense, this work meets important educational references on the subject, such as Seidman (2012) and Tinto (1993). In fact, our actions are based on the use of data mining for early identification of at-risk students as well as early, intensive, and continuous interventions to help them (Seidman's formula) (Seidman, 2012). In addition to helping these students in relation to the specific course, our work also aims at bringing them closer to the faculty and other students, facilitating their adaptation and integration into the academic environment as well as the construction of positive social interactions (Tinto's model) (Tinto, 1993).

3 Materials and methods

The data mining approach presented in this article includes the two major phases illustrated by Figure 1. The first phase, illustrated at the top of the figure, refers to the process of Knowledge Discovery in Databases (KDD), which is adopted to train several data mining algorithms to learn complex correlations among the features modeled to identify at-risk students. Here the input data are records of students from previous years, for which the results (failure or approval) are already known. Pre-processing strategies like normalization and the filling in missing values are also considered. Further, the dataset is used to train the predictive models and make them able to identify the students with high chance to fail. The next step involves the Post-processing of these models, which is directly related to the evaluation and interpretation of the assumptions built by the learned models which may result in knowledge at the end of this whole process.

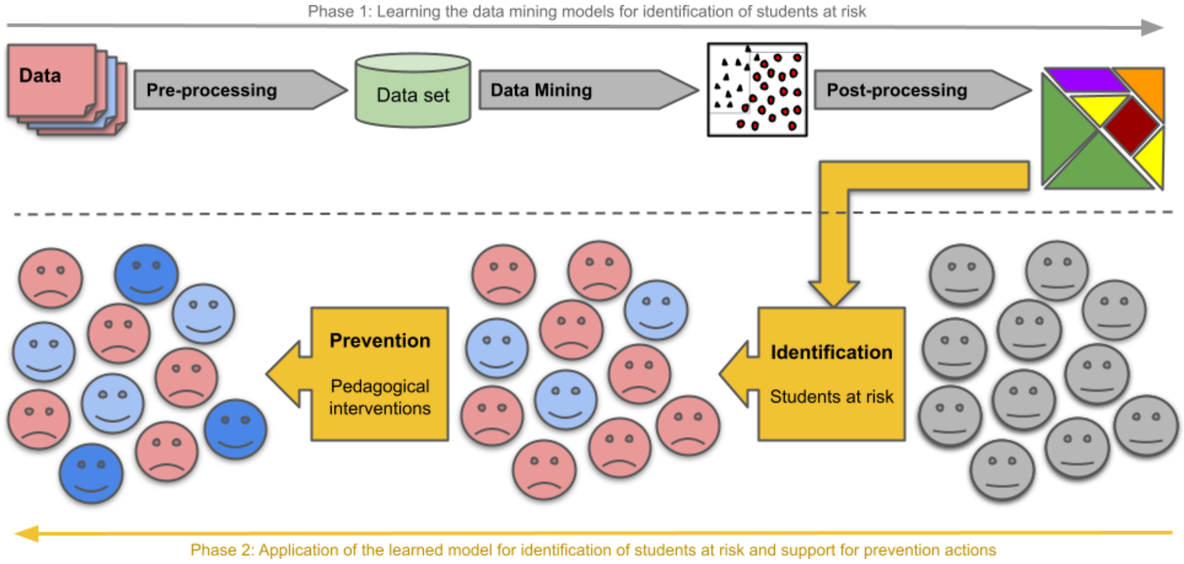

Phase 1: Learning the data mining models for identification of students at risk
Data
Pre-processing
Data set
Data Mining
Post-processing
Prevention
Pedagogical interventions
Identification
Students at risk
Phase 2: Application of the learned model for identification of students at risk and support for prevention actions

Figure 1: Flowchart of the methodology developed for the identification of students at risk of failure and retention prevention.

The second phase, represented at the bottom of Figure 1, comprises a real-world scenario where we apply the knowledge acquired in the previous phase (provided by the predictive models) to early identify students at risk of failure and also to guide several pedagogical interventions to prevent that. In a few words, we expect that students initially with a high chance to fail may have better chances to pass at the end of the semester when these interventions are employed. In the following we detail the steps of each phase.

3.1 Pre-processing

The dataset consists of data from students of the Bachelor's Degree in Information Systems at the Federal University of Uberlândia. It comprises 98 student records collected in accordance with the Research Ethics Committee of the Federal University of Uberlândia (CAAE 86758718.0.0000.5152), referring to students enrolled in the ICP course between 2015 and 2017. Each record is denoted by the features shown in Table 1. The target attribute (class) to be predicted is the student's situation at the end of the ICP course: passed or failed (both by reproving or dropping out). Notably, 69 samples belong to the "failed" class and 29 to the "passed" class, which indicates that the dataset is unbalanced.

Another concern related to the database is the missing values, i.e., features in which the student did not provide information. There are 59 missing values in the whole dataset, and a total of 47 records (and eight features) have at least one missing value. In total, features "Math average in high school" and "Distance (km) to the University" have 20 missing values each. How-

Table 1: Feature description of the dataset

Feature	Type	Values
Gender	Nominal	{1: male, 2: female}
Age (at the course admission)	Discrete	$\{17 \leq x \leq 29 \mid x \in \mathrm{Z}\}$
Race (by color skin)	Nominal	{1: white, 2: brown, 3: black, 4: yellow or indigenous}
Work situation	Nominal	{0: never worked, 1: unemployed, 2: employed}
Civil status	Nominal	{1: single, 2: married, 3: divorced}
Number of children	Discrete	$\{0 \leq x \leq 2 \mid x \in \mathrm{Z}\}$
High school type	Nominal	{1: public, 2: partially public, 3: private}
Math average in high school	Continuous	$\{50 \leq x \leq 92.73 \mid x \in \mathrm{R}\}$
Who do you live with	Nominal	{1: family, 2: relatives or friends, 3: alone, 4: student republic, 5: boarding house}
Distance (km) to the University	Continuous	$\{1.5 \leq x \leq 6.3 \mid x \in \mathrm{R}\}$
Distance (km) to the parents' house	Continuous	$\{0 \leq x \leq 4908 \mid x \in \mathrm{R}\}$
Parents' marital status	Nominal	{0: single mother, 1: married, 2: divorced, 3: widow}
Family is beneficiary of any social program	Nominal	{0: no, 1: yes}
Family per capita income (in BRL)	Continuous	$\{125 \leq x \leq 1856.35 \mid x \in \mathrm{R}\}$
University entrance exam type	Nominal	{1: national exam, 2: regional exam, 3: transfer}
Grade in ICP course	Nominal	{0: reproved, 1: approved}

ever, only four samples (students) of our whole database have the values of both features missing simultaneously. In addition, only two samples of our database present more than two missing values simultaneously. Thus, despite the challenge of dealing with some missing values, they do not compromise our work as their corresponds to less than 5% of our features' values (a total of 1470). Furthermore, we also evaluated an additional treatment to deal with such missing values in which we replaced the missing data with the average/mode values of that features, one of the strategies most used in the literature (García-Peña, Arciniegas-Alarcón, & Barbin, 2014). Then, in addition to the "Original Dataset" (with missing values), we generated a second dataset named "Transformed Dataset", in which we applied the imputation strategy mentioned above.

3.2 Data Mining and Post-processing

This article investigated several classification algorithms from the literature available in the Waikato Environment for Knowledge Analysis (WEKA) software (Frank, Hall, & Witten, 2016). We also designed an experimental study to set the parameters of the techniques as this can directly influence the model results. The algorithms and respective parameters evaluated in this work are:

- k-nearest neighbors algorithm (KNN) with the number of neighbors $k \in \{1,2,\dots,20\}$;
- Decision tree (J48 algorithm) with confidence factor varying as $\{0.05, 0.15, 0.25, 0.50, 0.75, 0.90, 0.95\}$;
- Random forest (RF) with the number of trees varying as $\{2^1, 2^2, \dots, 2^{10}\}$.
- Naive Bayes (NB), without parameters;
- Multilayer Perceptron (MLP), with two parameters: learning rate $\alpha \in \{0.01, 0.05, 0.1, 0.2, 0.3\}$ and number of neurons in the hidden layer $n \in \{10, 20, 50, 100, 500, 1000\}$;
- Pruning-based rule induction (JRIP algorithm), with the number of optimization steps as $\{2^0, 2^1, 2^2, \dots, 2^{11}\}$.

To evaluate the performance of the algorithms, we considered both accuracy and F1 measure. Moreover, the dataset was divided into two sets in the simulations: training and test, using the 10-fold cross-validation method.

3.3 Identification and support for at-risk students

As pointed out in section 2, several related studies do not consider the application of their predictive models in real-world situation neither present information on how these models might help to prevent failure or dropout. This section discusses such a point regarding our approach (Figure 1).

3.3.1 Identification of at-risk students

After completing the training phase of the data mining models, we apply the acquired knowledge to identify at-risk students considering real-world situations. Such situations consist of students who are taking the ICP course, i.e., their final result are unknown. If the previous phase aimed to induce a predictive model able to efficiently correlates the features and outputs of student records for which the outcome in the ICP course were known, this phase aims to (early) identify the students already taking the course with chances of failing.

3.3.2 Failure prevention using pedagogical interventions

Identifying at-risk students itself is not enough to prevent failure. However, this information is essential to understand probable causes related to the phenomenon and also to prepare and support practical actions to handle the problem. In this sense, this work also managed pedagogical interventions carried out by scholarship students and volunteers throughout the academic semester, under the planning and supervision of the project coordinator, and also with the support of some faculty members.

The use of supplementary content as an academic assistance program is a strategy that increases student performance through collaborative activities between teachers and students, and also among students (Arendale, 1994). Furthermore, literature research shows that mentoring (or tutoring) programs have been good allies in the fight against academic retention and dropout (Moschetti, Plunkett, Efrat, & Yomtov, 2018; Berger, 2019; Friedman et al., 2021).

In this sense, we present as follows the main actions carried out, which covered basic and advanced topics related to the ICP course:

- Explainable introductory talks presenting the high levels of retention and evasion in higher education, especially in the ICP course, as well as our data mining-based approach to prevent retention;
- Placement tests to analyze basic deficiencies in the students' formation;
- Intra-class support by tutors accompanying the course classes to support the lecturer and, mainly, establish a relationship of trust with the target students of the project;
- Extra-class support to help students, using email, instant messaging applications, and face-to-face meetings;
- Courses and mini-courses to cover topics of the course syllabus in which many students demonstrate difficulty;
- Programming contests to stimulate problem solving and motivate the study of logic and programming fundamentals;
- Programming task force to provide the students an intense period of study and immersion, covering all the topics in the course's syllabus;
- Coordinator meetings to analyze the effectiveness of our interventions and, at the same time, receive students' feedback regarding their improvements.

4 Experimental Results

Table 2 summarizes the results obtained by each algorithm described in the previous section. Analyzing their predictive performance in the Original and Transformed datasets, one can see that the former presented better results, probably because our generic imputation strategy generated some noise in the data.

The KNN algorithm obtained a regular result, and we emphasize the difference between the accuracy and the F1 score of the Transformed Base (difference of approximately 10%), caused by the imputation strategy that makes records that eventually should be distant in Euclidean space closer. The J48 algorithm also presented regular performance considering the accuracy but with considerable performance loss in the two databases when evaluated by the F1 measure. This difference occurred because, in these cases, the algorithm classified the vast majority of records as belonging to the same class, finding a relatively high accuracy but penalized by the F1 score, which considers the measures of precision and recall.

The RF algorithm uses a voting scheme among a committee of decision trees as a classification method. However, the fact that the database is unbalanced and with few records challenges

Table 2: Predictive results of the algorithms in terms of accuracy and F1 measure.

	Original Dataset		Transformed Dataset	
Algorithm	Accuracy(%)	F1(%)	Accuracy(%)	F1(%)
KNN	73.46 ± 1.78	72.10 ± 5.52	70.40 ± 0.71	60.60 ± 4.75
J48	70.40 ± 1.60	63.60 ± 7.37	72.44 ± 1.58	62.70 ± 5.60
RF	74.48 ± 3.40	68.90 ± 9.87	71.42 ± 3.53	68.20 ± 6.20
NB	69.38 ± 1.72	68.70 ± 5.00	57.14 ± 3.12	54.90 ± 5.39
MLP	75.51 ± 4.30	74.60 ± 3.87	62.24 ± 3.22	67.50 ± 3.62
JRIP	**81.63 ± 2.79**	**80.70 ± 5.47**	**80.61 ± 5.09**	**79.40 ± 1.31**

the technique's performance concerning the less represented classes. On the other hand, it is worth mentioning that the technique presented a slight performance variation over the two databases and acceptable predictive results. On the other hand, the NB algorithm obtained the worst results, with all metrics below 70%, probably because this algorithm does not explore any correlation between the attributes of the objects (independence assumption).

Considering now the performance of the MLP, we observed that it was more efficient in the Original Base than the Transformed Base. We conclude that the imputation strategy may difficult the algorithm's performance in this database, making it challenging to classify different objects by assigning them the same values for the missing data.

Finally, JRIP achieved the best results due to its ability to deal with unbalanced data. Furthermore, the algorithm is robust to noise. It employs a *divide-and-conquer* strategy to iteratively divide the training data and induce rules that maximize the number of classified samples using the sequential coverage technique (Cohen, 1995). In the results of Table 2, we observe the slight variation between accuracy and F1 score in the two datasets, a fact that few algorithms were able to achieve.

In order to examine a possible statistical difference between the performance of the algorithms, we adopted the *McNemar* statistical test (Dietterich, 1998). The test creates a contingency table from the predictions made by two classifiers ($\hat{f}_A$ and $\hat{f}_B$), as represented in Table 3. In this example, n_{11} denotes the number of objects that both classifiers predict correctly, n_{00} number of objects that both misclassified, n_{01} and n_{10} mean, respectively, the number of samples only $\hat{f}_B$ or $\hat{f}_A$ classify correctly. The null hypothesis states that two algorithms have the same error rate. Thus, the *McNemar* test compares the distribution of expected error counts (n_{01} and n_{10}) concerning the observed ones based on the χ^2 distribution under the null hypothesis (Dietterich, 1998).

Table 3: Contingency table of the *McNemar* test.

		$\hat{f}_A$	
		correct	incorrect
$\hat{f}_B$	correct	n_{11}	n_{01}
	incorrect	n_{10}	n_{00}

The statistical tests were applied considering the Original Base, as we achieved the best predictive performance in this data set, and the performance measure F1 for taking into account the imbalance of the data. Table 4 presents the p-value obtained by the *McNemar* test considering all pairs of algorithms. Considering the 95% confidence level (i.e., $\alpha = 0.05$), the following statements can be derived from the test: MLP has a statistically lower error rate than J48, and JRIP has an error rate statistically smaller than J48 and NB. As the database is small, we also performed analyses considering a 90% confidence level ($\alpha = 0.1$). We noticed that RF has a statistically lower error rate than J48 and that JRIP has a statistically lower error rate than RF.

Table 4: P-values of the *McNemar* test. Results addressing the confidence level of 90% ($\alpha = 0.1$) were boldfaced.

	KNN	J48	RF	NB	MLP
J48	0.1003				
RF	1.0000	**0.0543**			
NB	0.5562	0.3268	0.3827		
MLP	0.8312	**0.0310**	1.0000	0.3447	
JRIP	0.1530	**0.0005**	**0.0961**	**0.0190**	0.3613

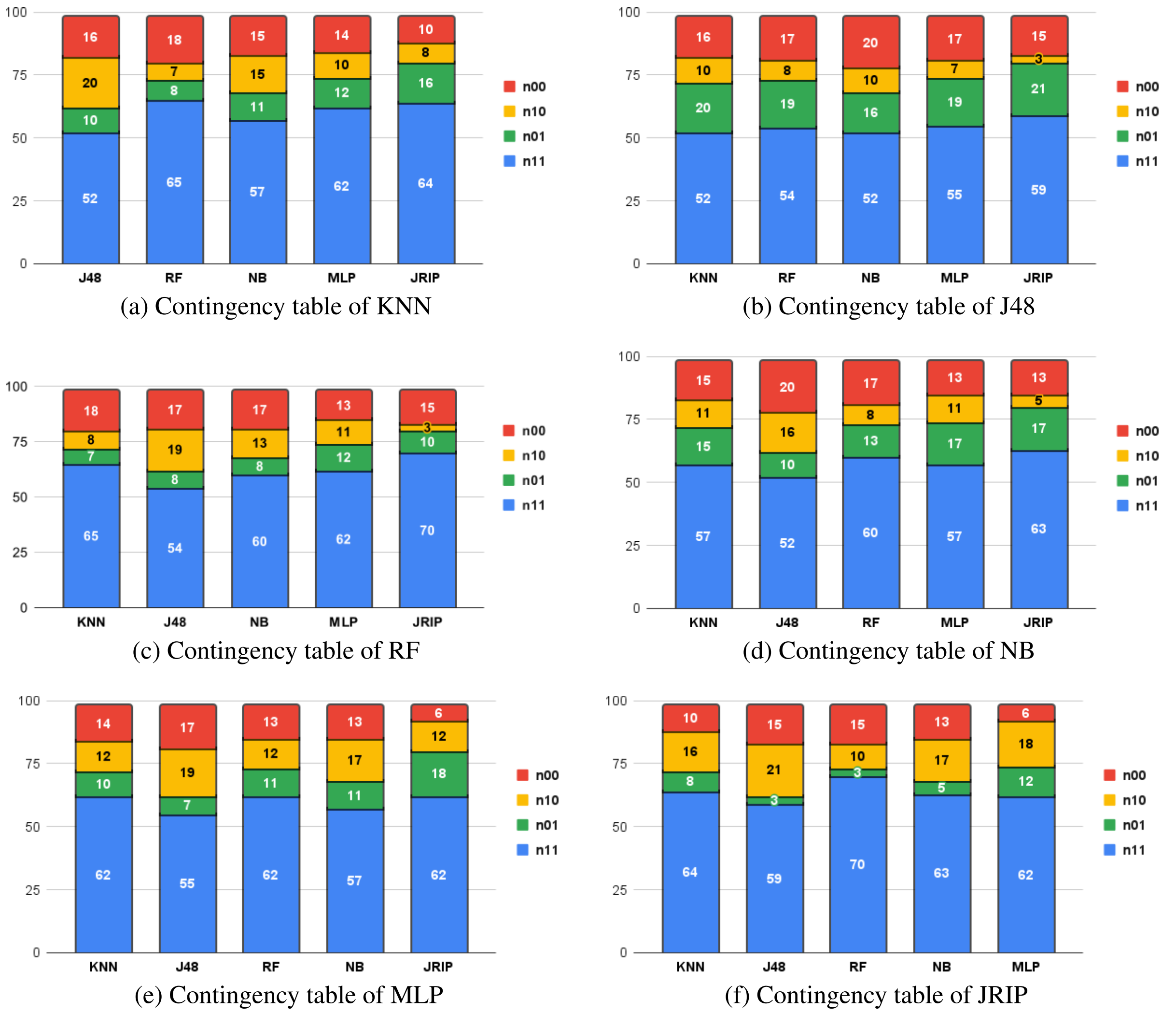
For a more detailed analysis of the statistical tests, Figure 2 presents the contingency tables generated by the *McNemar* tests between all pairs of algorithms, correlating their error rates. In Figure 2(b), it is possible to observe the poor performance of J48 against other algorithms, while Figure 2(f) highlights the good performance of JRIP. We also conclude that there is a potential for combining two or more algorithms, such as JRIP and MLP, which together missed only six objects in the database. In this sense, the investigation of ensembles strategies capable of efficiently combining the predictions of such algorithms may result in a significant performance improvement.

Finally, it is essential to mention that as the database has only a reasonable amount of samples, the statistical tests were adopted here more as a cautious and prudent analysis than as a definitive result. We believe that the path to a more coherent and accurate analysis is to add more objects in the database, which has been the target of several efforts by the coordinator of this project with the academic community, including the course students.

5 Results of prevention actions

This section presents the context in which the methodology presented in this article has been applied. It is important to emphasize that data mining itself does not solve the problem of failure and dropout. However, it can direct the efforts of managers, teachers, and even the students themselves towards more effective actions to combat the problem.

From the point of view of data mining, the rules obtained by the JRIP helped to identify at-risk students. For example, Table 5 presents the three rules induced by the model presented in the previous section, which provided the highest predictive performance among all the methods under analysis. In addition to guiding the coordinator's actions, these rules are essential for academic management. For example, rules 1 and 3 confront student aid rules commonly adopted by many higher education institutions, which do not usually provide housing aid for students who already reside in the same city as the campus. Now, according to the rules obtained by our model, low-



(a) Contingency table of KNN

(b) Contingency table of J48

(c) Contingency table of RF

(d) Contingency table of NB

(e) Contingency table of MLP

(f) Contingency table of JRIP

Figure 2: Summary of the contingency tables obtained for each classification algorithm.

income students who live far from the campus should be assisted by the institution to live closer.

Table 5: Example of rules obtained by the JRIP algorithm.

ID	**JRIP rules to identify at-risk students**
1	IF distance to the parents' house is less than *41km* AND family per capita income is less than *R$ 766* AND high school type is *public*
2	IF math average in high school was less than *83*
3	IF distance to the parents' house is less than *219km* AND distance to the university is greater than *2.9km*

In addition to identifying at-risk students, this work included a series of actions and interventions involving lectures, monitoring, assistance, courses, mini-courses, and programming competitions. These actions had the support of the faculty and were carried out by scholarship students and volunteers throughout the academic semester, under the planning and supervision of the coordinator. Figure 3 summarizes the percentage of approved, disapproved, and dropouts related to the course since the beginning of the degree in our campus. In 2018 and 2019, the two stages of our project were carried out: identifying at-risk students using data mining and actions/interventions to combat failure. Indeed, several factors may contribute to such an improvement like an eventual change of the teacher or student profile. However, the feedback received by those involved (teachers and students) suggests that the methodology presented in this article has some contribution in the increasing of the number of students approved in the course and also in the decreasing of the dropout percentages.

In order to analyze the statistical significance of such results, we conduct an unpaired (independent) two-tailed t-test to evaluate if there is any significant difference between the percentage means of approved students in the period in which our approach was applied (2018-2019) compared to the period it was not (2013-2017). The data considered are the same presented in Figure 3, which comes from sample sizes of 50 (2013), 76 (2014), 81 (2015), 64 (2016), 84 (2017), 79 (2018) and 73 (2019) students. Considering a significance level $\alpha = 0.1$, the null hypothesis is rejected (ρ-value = 0.09), which means that the percentage of students approved in each period are significantly different. Such a result gives pieces of evidence about the contributions of our approach to confront retention and evasion phenomenons.

6 Conclusion

This paper presented a methodology to address the student failure issue in academic in-class courses. Specifically, the objective was to identify which factors contribute to students failure in high failure rates courses, such as ICP ($\approx$ 78%), and to conduct a series of actions and interventions with these at-risk students in order to reverse the scenario. This investigation was based on a dataset of 98 students enrolled in an introductory programming course in a Bachelor's Degree in Information Systems at Federal University of Uberlândia, Brazil.

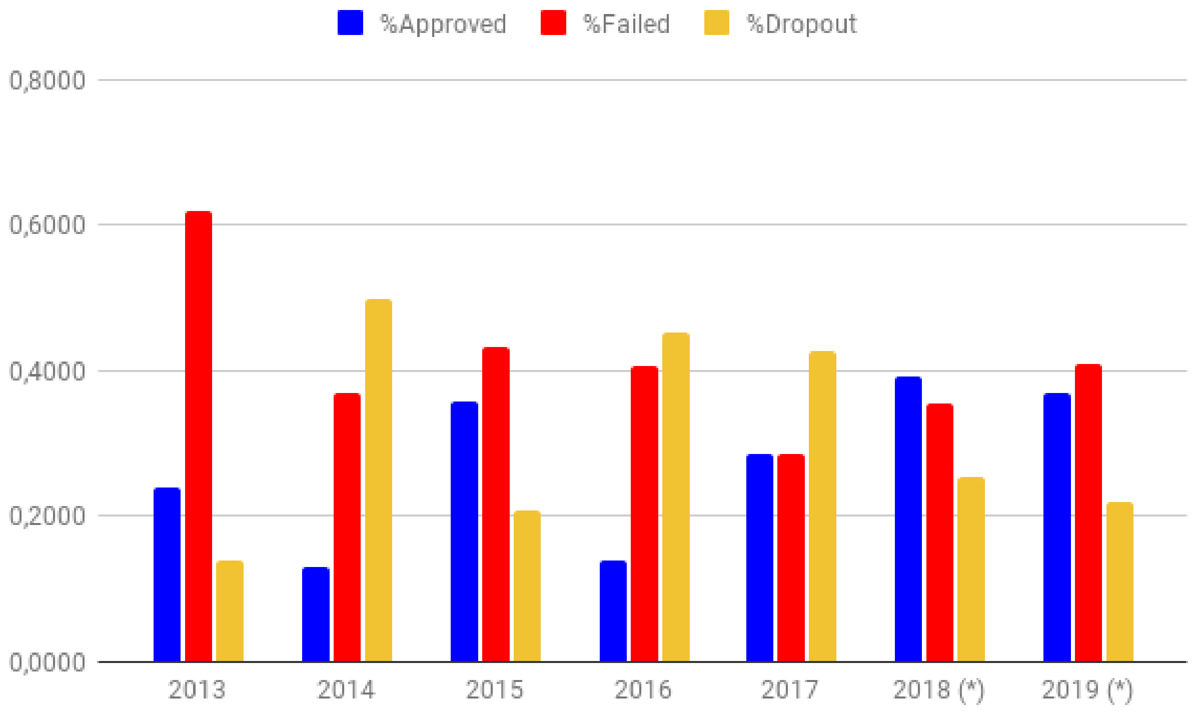


Figure 3: Historical percentage of students' results in the ICP course presented in this research from 2013 to 2019. (*) Years in which the methodology proposed in this work was applied.

We have investigated several classification algorithms from the literature in order to find out those that perform best in identifying students who are most likely to fail the course. However, only identifying at-risk students is not enough to prevent failure. Thus, we conducted several pedagogical interventions with the help of scholarship students and volunteers for those students identified by the predictive models as failure-prone students.

Regarding the two research questions that guided this research, we obtained the following answers:

RQ1. *Is it possible to develop a predictive model capable of revealing the factors that most contribute to the identification of students with a greater chance of failing the course?*

Answer RQ1. Yes, it was possible to build an accurate prediction model (above 80% in terms of accuracy and F1 score) to identify the factors that contribute to failure in the course. In our context, the distance from the housing to the parents' home, family's per capita income, type of school attended in high school, average in mathematics, and distance from home to the university were key factors for identifying at-risk students;

RQ2. *Can such factors help guide practical actions to combat failure in the course?*

Answer RQ2. Yes, the results of the intervention procedure guided by the identified factors contributed to a significant improvement in students' course performance, with an increase of the number of successful students and a decrease of the number of dropouts.

We also notice the lack of a standard methodology for works related to the subject, which may eventually reduce the relevance of discussions and findings presented by these studies. In this sense, another contribution of this work is a solid experimental procedure based on the application of well known statistical tests, which we expect to serve as a reference for other educational data mining works.

As future work, we intend to apply this methodology as an institutional policy to provide an effective tool to address students' failure and dropout in other courses at the Federal University of Uberlândia. Additionally, we expect to improve our model's predictive performance by investigating other data preparation methods and ensemble-based learning strategies. Finally, we intend to employ visual strategies in the obtained results, as well as in the original data, in order to identify relevant and strategic patterns that help educational managers and researchers to better comprehend the educational context and their specificity. We believe such strategies may better guide these experts to propose effective educational interventions and mitigate students failure.

Acknowledgement

Authors thank the students and faculty members that supported the project. They also thank the Institutional Program of Assisted Graduation (PROSSIGA) of the Federal University of Uberlândia by the financial support.

References

Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, *5*(2), e01250. doi: 10.1016/j.heliyon.2019.e01250 [GS Search]

Arendale, D. R. (1994). Understanding the supplemental instruction model. In D. C. Martin & D. R. Arendale (Eds.), *Supplemental Instruction: Increasing student achievement and retention. (New Directions in Teaching and Learning, No. 60, pp. 11-21).* San Francisco: Jossey-Bass. doi: 10.1002/tl.37219946004 [GS Search]

Baker, R., Isotani, S., & Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, *19*(02), 03–13. doi: 10.5753/rbie.2011.19.02.03 [GS Search]

Beltran, C. A. R., Xavier-Júnior, J. C., Barreto, C. A., & Oliveira Neto, C. (2019). Plataforma de aprendizado de máquina para detecção e monitoramento de alunos com risco de evasão. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)* (Vol. 30, pp. 1591–1600). doi: 10.5753/cbie.sbie.2019.1591 [GS Search]

Berger, R. (2019). Dropouts thoughts on whether having a mentor would have helped them remain in school. *Open Access Library Journal*, *6*. doi: 10.4236/oalib.1105718 [GS Search]

Cambruzzi, W. L., Rigo, S. J., & Barbosa, J. L. (2015). Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach. *Journal of Universal Computer Science*, *21*(1), 23–47. doi: 10.3217/jucs-021-01-0023 [GS Search]

Carrano, D., de Albergaria, E. T., Infante, C., & Rocha, L. (2019). Combinando técnicas de mineração de dados para melhorar a detecção de indicadores de evasão universitária. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)* (Vol. 30, pp. 1321–1330). doi: 10.5753/cbie.sbie.2019.1321 [GS Search]

Cohen, W. W. (1995). Fast effective rule induction. In *Machine learning proceedings 1995* (pp. 115–123). Elsevier. doi: 10.1016/B978-1-55860-377-6.50023-2 [GS Search]

Damasceno, I. L., & Carneiro, M. G. (2018). Panorama da evasão no curso de sistemas de informação da Universidade Federal de Uberlândia: Um estudo preliminar. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)* (Vol. 29, pp. 1766–1770). doi: 10.5753/cbie.sbie.2018.1766 [GS Search]

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, *10*(7), 1895–1923. doi: 10.1162/089976698300017197 [GS Search]

Frank, E., Hall, M. A., & Witten, I. H. (2016). The WEKA workbench. In *Data mining: Practical machine learning tools and techniques* (4th ed., chap. Online Appendix). Burlington, MA: Morgan Kaufmann. [GS Search]

Friedman, D. B., Yelton, B., Corwin, S. J., Hardin, J. W., Ingram, L. A., Torres-McGehee, T. M., & Alberg, A. J. (2021). Value of peer mentorship for equity in higher education leadership: a school of public health focus with implications for all academic administrators. *Mentoring & Tutoring: Partnership in Learning*, *29*(5), 500–521. doi: 10.1080/13611267.2021.1986795 [GS Search]

García-Peña, M., Arciniegas-Alarcón, S., & Barbin, D. (2014). Climate data imputation using the singular value decomposition: an empirical comparison. *Revista Brasileira de Meteorologia*, *29*(4), 527–536. doi: 10.1590/0102-778620130005 [GS Search]

Gottardo, E., Kaestner, C., & Noronha, R. V. (2012). Previsão de desempenho de estudantes em cursos EAD utilizando mineração de dados: uma estratégia baseada em séries temporais. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)* (Vol. 23). [GS Search]

Hawlitschek, A., Köppen, V., Dietrich, A., & Zug, S. (2019). Drop-out in programming courses – prediction and prevention. *Journal of Applied Research in Higher Education*, *12*(1). doi: 10.1108/JARHE-02-2019-0035 [GS Search]

Horton, D., & Craig, M. (2015). Drop, fail, pass, continue: Persistence in CS1 and beyond in traditional and inverted delivery. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education* (p. 235–240). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2676723.2677273 doi: 10.1145/2676723.2677273 [GS Search]

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, *1*(1), 6–47. doi: 10.18608/jla.2014.11.3 [GS Search]

Kampff, A. J. C., Ferreira, V. H., Reategui, E. B., & Lima, J. V. d. (2014). Identificação de perfis de evasão e mau desempenho para geração de alertas num contexto de educação a distância. *RELATEC: Revista Latinoamericana de Tecnología Educativa*, *13*(2), 61–76. Retrieved from http://hdl.handle.net/11162/134334 [GS Search]

Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *International conference on knowledge-based and intelligent information and engineering systems* (pp. 267–274). doi:

10.1007/978-3-540-45226-3_37 [GS Search]

Li, H., Ding, W., & Liu, Z. (2020). Identifying at-risk K-12 students in multimodal online environments: a machine learning approach. In *Proceedings of the 13th international conference on educational data mining (edm 2020).* [GS Search]

Manhães, L. M. B., Da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., & Zimbrão, G. (2012). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)* (Vol. 1). [GS Search]

Moschetti, R. V., Plunkett, S. W., Efrat, R., & Yomtov, D. (2018). Peer mentoring as social capital for latina/o college students at a hispanic-serving institution. *Journal of Hispanic Higher Education*, *17*(4), 375–392. doi: 10.1177/1538192717702949 [GS Search]

Neves, F., Campos, F., Dantas, M., David, J. M., Braga, R., & Stroele, V. (2021). Uso de aprendizado de máquina para detecção de risco de evasão no curso de licenciatura em computação. *Lynx*, *1*(2). Retrieved from https://periodicos.ufjf.br/index.php/lynx/article/view/35552 [GS Search]

Noetzold, E., & de L. Pertile, S. (2021). Análise e predição de evasão dos alunos de um curso de graduação em sistemas de informação por meio da mineração de dados educacionais. *RENOTE - Revista Novas Tecnologias na Educação*, *19*(1). doi: 10.22456/1679-1916.118525 [GS Search]

Oliveira, J. L., Ambrósio, A. P., Silva, U., Brancher, J., & Franco, J. J. (2020). Undergraduate students' effectiveness in an institution with high dropout index. In *2020 IEEE Frontiers in Education Conference (FIE)* (p. 1-7). doi: 10.1109/FIE44824.2020.9274108 [GS Search]

Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in chile. *Entropy*, *23*(4). doi: 10.3390/e23040485 [GS Search]

Pappas, I. O., Giannakos, M. N., & Jaccheri, L. (2016). Investigating factors influencing students' intention to dropout computer science studies. In *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education* (p. 198–203). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2899415.2899455 [GS Search]

Petersen, A., Craig, M., Campbell, J., & Tafliovich, A. (2016). Revisiting why students drop CS1. In *Proceedings of the 16th Koli Calling International Conference on Computing Education Research* (p. 71–80). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2999541.2999552 [GS Search]

Pimentel, T., Passos, C., Fernandes, I., & Goldschmidt, R. (2019). Mineração de padrões sequenciais de sentimentos: Um estudo de caso na detecção de propensão à evasão escolar na educação superior. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)* (Vol. 30, pp. 1411–1420). doi: 10.5753/cbie.sbie.2019.1411 [GS Search]

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1355. doi: 10.1002/widm.1355 [GS Search]

Salloum, S. A., Alshurideh, M., Elnagar, A., & Shaalan, K. (2020). Mining in educational data: Review and future directions. In *Aicv* (pp. 92–102). doi: 10.1007/978-3-030-44289-7_9

[GS Search]

Santos, K. J. O., Menezes, A. G., de Carvalho, A. B., & Montesco, C. A. E. (2019). Supervised learning in the context of educational data mining to avoid university students dropout. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)* (Vol. 2161-377X, p. 207-208). doi: 10.1109/ICALT.2019.00068 [GS Search]

Santos, R., Pitangui, C., Vivas, A., & Assis, L. (2016). Análise de trabalhos sobre a aplicaçao de técnicas de mineraçao de dados educacionais na previsao de desempenho acadêmico. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação* (Vol. 5, pp. 960–969). doi: 10.5753/cbie.wcbie.2016.960 [GS Search]

Santos, V., Saraiva, D., & Oliveira, C. (2021). Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação* (pp. 1196–1210). Porto Alegre, RS, Brasil: SBC. doi: 10.5753/sbie.2021.218167 [GS Search]

Seidman, A. (2012). Taking action. *A Retention Formula and Model for Student Success. In A. Seidman (Ed.), College Student Retention. Formula for Student Success*, 267–284. [GS Search]

Silva, R., Borges, B., Ferreira, M. d. F., Santos, I., & Andrade, R. (2021). Evasão em computação na UFC sob a perspectiva dos alunos. In *Anais do XXIX Workshop sobre Educação em Computação* (pp. 338–347). Porto Alegre, RS, Brasil: SBC. doi: 10.5753/wei.2021.15925 [GS Search]

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition*. ERIC. [GS Search]

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, *8*(2), 307–325. doi: 10.6339/JDS.2010.08(2).574 [GS Search]