

Educational Data Mining for Dropout Prediction: an Experience at a University in Southern Brazil

Mineração de Dados Educacionais para a Predição de Evasão: Experiência em uma Universidade do Sul do Brasil

Piero Salaberri
Sandra Dutra Piovesan
Valesca Brasil Irala

ABSTRACT


Dropout is a problem that plagues public and private higher education institutions around the world and strategies for analyzing the reasons for the phenomenon abound in scientific publications. Many works that aim to find the most appropriate and effective techniques and practices for identifying dropout inducers in students end up being based on the use of technologies to improve data analysis and achieve a greater volume of processed information. The present study aims to identify good practices for the use of data mining for educational information. For this purpose, existing practices in the literature were investigated for structuring research with data from a public university in the interior of the state of Rio Grande do Sul. The study includes practical tests with the Decision Tree algorithms C4.5, Random Forest and Neural Networks in different datasets. The work demonstrates that the Random Forest algorithm was able to be more accurate in identifying students at risk of dropping out. From this experience other institutions will be able to base themselves for the definition of their best practices.


Keywords: Dropout. College. Higher Education. Educational Data Mining. Algorithms.


Submitted: 17/08/23
Accepted: 25/04/24

RESUMO

A evasão é um problema que assola instituições de ensino superior públicas e privadas em todo o mundo e estratégias para análise dos motivos para o fenômeno abundam em publicações científicas. Muitos trabalhos que objetivam encontrar as técnicas e práticas mais apropriadas e efetivas para a identificação dos indutores de desistência no aluno acabam por se alicerçar no uso de tecnologias para aprimorar a análise dos dados e atingir um maior volume de informação processada. O presente estudo visa identificar boas práticas para o uso de mineração de dados para informações de cunho educacional.

Piero Salaberri 
pierosalaberri@unipampa.edu.br
Master's Degree
Universidade Federal do Pampa
Bagé / RS – Brazil

Sandra Dutra Piovesan 
sandrapiovesan@unipampa.edu.br
Doctorate Degree
Universidade Federal do Pampa
Bagé / RS – Brazil

Valesca Brasil Irala 
valescairala@unipampa.edu.br
Doctorate Degree
Universidade Federal do Pampa
Bagé / RS – Brazil

RESUMO

Para tanto, investigou-se práticas já existentes na literatura para a estruturação de uma pesquisa com dados de uma universidade pública no interior do estado do Rio Grande do Sul. O estudo conta com testes práticos com os algoritmos Árvore de Decisão C4.5, *Random Forest* e Redes Neurais em diferentes conjuntos de dados. O trabalho demonstra que o algoritmo *Random Forest* conseguiu ter maior precisão na identificação dos alunos em risco de evasão. A partir desta experiência outras instituições poderão basear-se para a definição de suas melhores práticas.

Palavras-chave: Evasão. Universidade. Ensino Superior. Mineração de Dados Educacionais. Algoritmos.

Introduction

The challenge of dropout rates in higher education is a global concern, impacting not only educational institutions in Brazil but also those worldwide. Extensive research exists in the scholarly literature, focusing on examining the underlying factors contributing to this phenomenon and devising effective strategies to mitigate it. A recurring theme in these studies is the recognition that the most successful interventions for retaining at-risk students are grounded in an understanding of the challenges they face. It is imperative that the insights gleaned from such analyses accurately reflect the realities of these students, serving as the foundation for targeted interventions to prevent dropout (Howlett, Ramesh, & Perl, 2013). From this understanding, it becomes possible to develop the necessary strategies to address this problem more effectively.

According to data presented by the Map of Higher Education in Brazil in 2023, only 26% of university students who entered universities in the year 2017 managed to complete their respective degrees. Throughout 2021, the dropout rate in the university reached 55%, contrasting with the 26% who graduated and the 18% who remained dedicated to their studies. This reality is a cause for concern, as it compromises the academic education of the students and results in the underutilization of resources allocated to education (SEMESP, 2023).

Furthermore, under analysis from the same publication, it is possible to identify that a larger contingent than 50% of students who entered higher education in 2017, and who should have completed their studies, did not manage to finish their

courses. During the analysis period, only 18% of students remained enrolled in universities (SEMESP, 2023).

In the quest to understand the underlying causes of school dropout, various analytical approaches are available. Those structured around computational techniques are deepened in this study, as they can expand the scope of data explored through statistical analyses, allowing the identification of patterns in which data can be grouped, correlated, or classified, according to the aspect under analysis.

By basing itself on data from student records at a university, the analysis of such data offers the opportunity to create knowledge in a way that anticipates possible behavioral trajectories of the students themselves. This can be achieved through experiments that systematically explore data, identifying relationships between variables that influence dropout. By adopting this approach, it is possible to develop predictive models that contribute to a deeper understanding of the factors that lead to school dropout, thus enabling the implementation of preventive and intervention measures. Data Mining constitutes an interdisciplinary domain, which essentially incorporates knowledge of statistical data analysis, machine learning, pattern identification, and representative information visualization (Cabena, Hadjinian, Stadler, Verhees & Zanasi, 1998) . This is largely based on the selection of available algorithms and the use of collected data sets for the analyses in question.

Establishing well-defined objectives is essential for obtaining relevant knowledge. In the process of knowledge discovery, these objectives are formulated based on the purposes of system utilization and can be categorized into two main approaches: verification or discovery (Fayyad, Piatetsky-Shapiro, & Smyth (1996). The focus of this study is on the search for the identification of the best algorithms for discovering patterns that can predict students at risk of dropping out. To this end, the study draws on the search for other experiences, also contextualized with data from higher education students.

Researchers like Souza (2021) indicate that decision tree and random forest algorithms efficiently analyze student data. However, various techniques need to be explored to identify the best option given the dataset and research objectives. This study, through the identification of algorithms with the highest potential and practical tests of those selected, aims to determine which options are most advantageous for the specific datasets and attributes analyzed during an investigation conducted

at a public university located in the interior of the state of Rio Grande do Sul. Such findings may support future researchers and institutions in making new parameterizations or approaches based on the considerations presented.

Dropout and Data Mining

Among the strategies available to analyze students and their entire context, data mining is one of the techniques that has been explored in various studies, as it can aggregate a large volume of data and provide knowledge discovery by performing association, classification, or clustering of data as per the researcher's objectives.

According to Han, Pei, and Kamber (2012), the Knowledge Discovery in Databases (KDD) process is a sequence of steps aimed at extracting knowledge from information contained in large databases. It is crucial to emphasize that KDD is an iterative process, implying that the execution of the steps does not follow a straight line from start to finish. It is often necessary to return to previous stages of the process to proceed forward again (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) .

Fayyad, Piatetsky-Shapiro and Smyth (1996) also propose dividing the Knowledge Discovery in Databases process into six distinct stages:

- **Data Preparation:** In this stage, the focus is on incorporating relevant information into the application and establishing the goals to be achieved by the process.
- **Data Cleaning:** The objective here is to eliminate data that may distort the analysis. It involves applying strategies to remove noise, deal with missing values, and even transform variables to reduce complexity, aiming to improve the performance of the analysis algorithms.
- **Data Selection:** At this stage, the decision is made on which data set or subset will be the target of the analysis process.
- **Data Mining:** At this point, the most appropriate data mining task to achieve the process objectives is chosen, as well as the most suitable technique for the task.
- **Incorporation of Prior Knowledge:** This stage involves interpreting the discovered model, evaluating its accuracy, and seeking improvements. This

allows returning to any earlier phase of the process, eliminating redundant or irrelevant patterns.

- **Interpretation of Results:** Here, the obtained results are integrated into the system, allowing actions based on the acquired knowledge. This may involve informed decision-making or documentation and presentation of the results to stakeholders.

Specifically, the Data Mining stage utilizes artificial intelligence techniques to identify similarities or discrepancies between data. The goal is to automatically discover patterns, anomalies, and rules, transforming apparently hidden data into valuable information for decision-making or result evaluation.

Baker, De Carvalho, Da Costa and Neves (2011) highlight that among the available techniques, the main ones used in an educational context can be categorized as:

- **Prediction:** seeks to predict the value of a specific attribute either by classification (identifying which class a record belongs to or what behavior the chosen attribute behaves for each subset of data (e.g., student)) or by regression (predicting a numeric attribute based on a data set).
- **Clustering:** aims to identify and group similar records.
- **Association:** aims to discover connections between variables in a data set.

According to Fayyad, Piatetsky-Shapiro and Smyth (1996), the purpose of predictive tasks is to anticipate the value of a specific attribute, also called a variable, based on the values of other attributes. At a more conceptual level, prediction uses a combination of variables to estimate other variables or unknown values. For example, by collecting data on students enrolled in higher education institutions, it is possible to examine all available data related to students' academic performance and, through data mining techniques, identify characteristics associated with dropout. Using the acquired knowledge, it is feasible to identify new students who may also be at risk of dropping out.

Educational data mining has proven to be a promising tool for addressing university dropout. This approach uses data analysis techniques to identify patterns and trends in student behaviors, allowing for a better understanding of the causes of dropout and enabling the adoption of preventive measures.

EDUCATIONAL DATA MINING

Educational Data Mining emerged as a research field that applies data mining techniques to educational data sets. Its purpose is to deepen the understanding of how students learn and interact with the educational environment so that such knowledge discovery can improve the educational outcomes of the students themselves.

Educational systems host a large volume of information instances from various systems and records, containing different data formats and dimensions (Romero & Ventura, 2013). These techniques also play a crucial role in the behavioral investigation of students, thus assisting in the planning, development, and operationalization of more effective technological solutions to support both students and managers and educators.

When we employ such techniques for prediction, the aim is to develop models that identify elements linked to the data. This is achieved by analyzing and combining the inherent characteristics of the data being examined, which are referred to as predictor variables. Obtaining information about students should seek to extrapolate the investigation of academic records, exploring options that reflect the most about the student's life, even if it requires the application of questionnaires or opinion polls (Baker, De Carvalho, Da Costa, & Neves, 2011). Educational Data Mining can play different roles depending on how the research is structured, from analyzing academic performance to predicting potential at-risk students, so that the institution itself can provide support and induce student retention (Souza, 2021).

After the data collection and preparation stage, we move on to the experimental testing phase. For this, the choice of algorithms and the tool to be used for their execution are necessary. Among the tools available for use are Weka, Rapidminer, KNINE, Orange, R Studio, among others. Among the main classification techniques are different types of classifiers, such as those based on decision trees, rule-based, neural networks, support vector machines, and naive Bayes classifiers. Each approach employs a learning algorithm to discern a model that fits more accurately the relationship between the set of attributes and the class labels of the input data. The resulting model, generated by the learning algorithm, must effectively adjust to the input data and be able to accurately predict the class labels of records that have not yet been observed (Brandão, 2018). As examples of algorithms based on the clas-

sification logic, we have Decision Trees C4.5, CART, and Random Forest; Multilayer perceptron Artificial Neural Networks; Naive Bayes; Linear and Logistic Regression; and Support Vector Machines (SVM), for instance.

The next section discusses the tests performed, as well as the process of choosing algorithmic solutions and tools for execution, indicating performance metrics for each subset of data used.

Materials and Methods

In general, decision tree-based algorithms demonstrate remarkable performance, consistently ranking among the leaders or standing out in comparisons, as evidenced in studies conducted by Fernandez-Garcia, Gonzalez, Rico and Prieto (2021), Franco, Martínez, and Domínguez (2021), and Lee and Chung (2019). Furthermore, investigations comparing different iterations of the decision tree algorithm have been identified. Sunday, Jekayinoluwa, Adedokun, Ajao and Yusuff (2020) and Hamoud, Hashim, and Awadh (2018) selected the C4.5 algorithm as the superior option in terms of overall performance when contrasted with the ID3 variants and, in the case of the former work, with the Random Tree and REPTree variations. Niy-oGiSubizo, Nduwimana, Nzobonimpa and Nkurunziza (2022) opted for the Extreme Gradient Boosting version.

Among all the analyses conducted, the algorithm that stood out as the most effective in studies by Alboaneen, Alenezi, Alrumayh, Almutairi, Alanazi, and Alotaibi (2022), Flores, Heras, and Julian (2022), Palacios, Arenas, Jimenez and Villanueva (2021), Perez-Gutierrez (2020), Yağci (2022), Al-Fairouz, and Al-Hagery (2020). Notably, in several instances, as exemplified by Kabathova and Drlik (2021), the Random Forest algorithm showed significantly superior performance compared to other evaluated algorithms, exhibiting percentage differences in accuracy metrics of up to almost 30%. In the context of the study by Fernandez-Garcia, Gonzalez, Rico and Prieto (2021), the algorithm demonstrated better performance for students in the 3rd and 4th semesters, where there is a considerable availability of academic records.

Other works such as those of Nabil, Seyam, and Abou-Elfetouh (2021), Nuan-kaew, Namahoot and Phewchean (2020), Siddique, Alam and Marwah (2021), and

Miranda and Guzmán (2017) observed that algorithms based on neural networks achieved the best performance in the analyzed data groups. On the other hand, in the study conducted by Yağci (2022), the performance was notably similar to that of Random Forest, which stood out as the best-performing option. There are still works like that of Nascimento, de Carvalho, Lima, Neves and Freitas (2023), which identified the Naive Bayes algorithm as the most tested in data mining studies. However, few are those in which this algorithm performs the best.

After evaluating the mentioned studies, from the perspective of a study aiming to predict the probability of students dropping out of higher education, using mainly demographic and academic information, the most promising algorithms were found to be: C4.5 Decision Tree, Random Forest, and Neural Networks. Therefore, the study will use these algorithms to determine which one presents the best performance and ability to extract the most relevant information in the various analysis situations that will be developed.

Cross-validation technique is widely employed to ensure comprehensive use of the dataset, dividing it into several parts, also known as folds. Cross-validation with $k = 10$ folds is often cited in the literature as the best value to be used (Castro & Ferrari, 2006). Thus, all model approaches will undergo 10-fold cross-validation, to evaluate whether this approach affects performance, considering that this technique has proven effective in the substantial majority of examined studies.

The models were developed using the Weka platform (Waikato, 1999), which offers an interface for model creation, adjustment, and evaluation, as well as for data submitted to testing. Alturki, Hulpuş, and Stuckenschmidt (2020), in their review, emphasized that this tool was the most commonly used in research dedicated to predicting dropout in higher education over the past decade. Authors like Nascimento, de Carvalho, Lima, Neves and Freitas (2023) corroborate this finding, indicating that the WEKA tool is among the most frequently used in studies aimed at investigating the best approaches to institutional data exploration. Martins (2017) concluded that WEKA offers a simpler interface, resulting in a more practical tool, with better usability, and a shorter learning curve. Additionally, it was observed that it has a satisfactory amount of documentation.

As a data source for the research experiment, the author requested access to the institution's database to obtain information related to the students. The institu-

tion's database has around 40,000 links between regular students, with completed links, and dropouts. After obtaining authorization, the author gained full access to the records stored in the databases maintained by the technical team responsible for data management.

In this context, a variety of information was available for access and analysis, covering demographic data (data related to city of origin, age, gender, and marital status); enrollment data (data on type of access and average used for admission); course-related data (distance learning or face-to-face course type, bachelor's/teaching/technological degree, and shift); personal/family order data (ethnicity, whether the student has a disability and what it is, public or private school of origin, number of family members, and income data); academic link data (total number of passed and failed subjects, number of subjects passed and failed per semester from 1 to 10, required and passed workload for course completion, and workload of required and passed complementary activities); student assistance (benefit requests submitted and granted, types of benefits received with number of installments received and average value); library usage (data on loans of physical and digital books); and University Restaurant (number of meals, amount paid by the student, and amount subsidized by the institution). This comprehensive set of information allowed for an investigation of the possible implications of each of these attributes in the context of student dropout at that university.

Studies conducted by Lanot (2012) and Borin (2014), set in the largest academic unit of the university, highlight the importance of considering the temporal context of approvals and failures about the curricular structure of the courses. It is essential to understand in which phase of the curriculum these events occur. However, analyzing this dynamic for each course in the institution would become impractical due to the time required for execution. To circumvent this limitation, the collected data were categorized by semester in which the discipline is integrated into the course. Thus, it was possible to examine an approach that also provides a consistent view compared to the models generated individually for each course.

To understand the metrics used in the next section of the article, titled Data Analysis, it is necessary to express the definition of each indicator present. All metrics stem from the confusion matrix, which is built by the algorithm itself after its execution. According to Castro and Ferrari (2006), the table in question represents

the model's hits and misses when compared to the desired outcome. The model's hits form the main diagonal, while the other values correspond to the errors that occurred. When focusing on a specific class, which has two possible values (for example, "student dropped out: YES or NO"), this is called a binary scenario. In this context, the most common representation is exemplified by the matrix shown in Figure 1.

Figure 1. Confusion matrix.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Source: (Castro & Ferrari, 2006).

The matrix fields are filled with the following components:

- **True Positive (TP):** indicates the correct classification of the positive class;
- **False Negative (FN):** occurs when the model predicts the negative class, but the actual value is positive;
- **False Positive (FP):** occurs when the model predicts the positive class, but the actual value is negative; and
- **True Negative (TN):** represents the correct classification of the negative class.

Based on the values of the fields described above, performance metrics can be calculated as follows:

- **Accuracy:** This is the number of correct classifications divided by the total number of classifications. It can be expressed by the formula $(TP+TN)/(TP+FP+TN+FN)$;

- **Precision:** This measures the accuracy of the algorithm, i.e., among all the positive class predictions made by the model, how many are correct. When False Positives have a greater impact than False Negatives, precision can be a relevant metric for analysis. It can be expressed as $TP/(FP+TP)$;
- **Recall:** This evaluates how many of the expected positive cases were correctly classified. It is an interesting metric when False Negatives are considered more harmful than False Positives. It can be expressed as $TP/(FN+TP)$; and
- **F-Score:** This is the harmonic mean between precision and recall, i.e., when a low F1-Score is observed, it indicates that either precision or recall is low. It can be expressed as $(2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$.

Data Analysis

Initial tests were conducted using a set of 86 attributes. The results of these tests are presented in Table 1. Initially, the algorithms Random Forest (RF), Decision Tree C4.5, and Neural Networks were used with the default configurations of the WEKA software. It was observed that the C4.5 and RF algorithms demonstrated the best performances, achieving accuracy rates of 61.25% and 60.07%, respectively. However, when analyzing the Confusion Matrix, it was identified that the RF algorithm achieved a higher number of correct predictions for students who dropped out (3428 correct predictions compared to 3286 by the C4.5 algorithm). Given a context where it is crucial to provide support to at-risk students, the decision was made in favor of the algorithm with higher precision in detecting these students.

Table 1. Metrics of the algorithms tested with 86 attributes.

Algorithm/Dataset	Metric	Value	Confusion matrix (a= YES, b= NO)									
<i>Random Forest on original dataset (86 attributes)</i>	Accuracy	60,0658%	<table border="1"><tr><td></td><td>a</td><td>b</td></tr><tr><td></td><td>3428</td><td>3692</td></tr><tr><td></td><td>2743</td><td>6251</td></tr></table>		a	b		3428	3692		2743	6251
		a		b								
		3428		3692								
		2743		6251								
Precision	0,596											
Recall	0,601											
	F-Score	0,596										

C4.5 on original dataset (86 attributes)	Accuracy	61,2511%	a	b
	Precision	0,608	3286	3834
	Recall	0,613	2410	6584
	F-Score	0,605		
Redes Neurais Multilayer Perceptron on original dataset (86 attributes)	Accuracy	59,3583%	a	b
	Precision	0,588	3164	3956
	Recall	0,594	2593	6401
	F-Score	0,586		

Source: Authors (2023).

By examining the confusion matrices of the algorithms evaluated with the set of 86 attributes, it is observed that the overall accuracy is low, ranging from 59.36% to 61.25%. The Neural Networks algorithm ranked last in this evaluation, requiring higher computational capacity and consequently more execution time. Due to its lower effectiveness, it was excluded from the next testing stages. Furthermore, a higher number of false positives may indicate more students with profiles similar to those identified as at risk of dropping out, serving as an alert for which students the university should direct its attention. This reinforces Random Forest as the best option among the tested group.

The field of machine learning not only allows for evaluating the effectiveness and accuracy of algorithms but also enables investigating the contribution of each attribute throughout the process. This analysis provides insights into the relevance of the decisions made by the algorithm based on the information available in the database.

In decision tree algorithms, the importance or relevance of attributes during the training of a data set is calculated by selecting the variables or features at each node that maximize error reduction in the overall prediction process. In this context, the most significant attributes in a decision tree are ranked based on their ability to reduce error when considered, weighing this reduction according to the number of observations related to the node. In the case of the Random Forest algorithm, which involves randomly splitting the original dataset and creating several smaller trees, this procedure is performed individually for each of the trees created. Subsequently, these individual importances are aggregated through averages to determine the overall importance of a specific feature (Thorn, 2020).

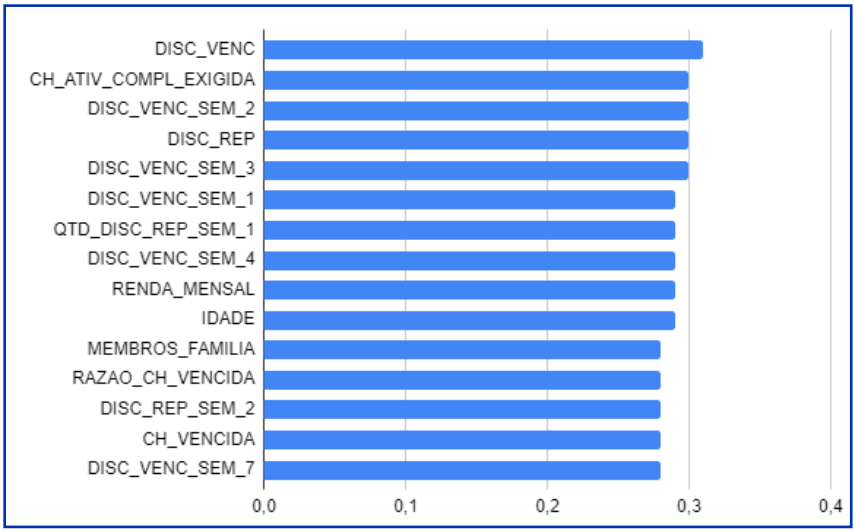
The initial testing phase revealed a considerable number of attributes with low relevance in the data mining process. The relevance of the attributes in the analyzed dataset is demonstrated in Figures 2 and 3. In Figure 2, the 15 most relevant attributes are listed, with index values ranging from 0.31 to 0.28. From this selection, it was notable that the subjects completed by students emerged as the most significant attribute, indicating its predominant relevance. The attributes related to completed subjects (DISC_VENC, DISC_VENC_SEM_2, DISC_VENC_SEM_3, DISC_VENC_SEM_1, DISC_VENC_SEM_4, CH_VENCIDA, DISC_VENC_SEM_7) suggest that academic performance plays a fundamental role in the decision-making process regarding dropout, as students who complete more subjects and have fewer failures generally show a reduced risk of leaving their studies. This observation is consistently aligned with publications aimed at dropout analysis, which point out that academic data, especially related to the number of approved and failed subjects, represent the strongest indicators for predicting dropout.

The second most important attribute is the number of failed subjects. Additionally, it is noteworthy that subjects completed in the first two years of the educational journey have high significance, while the number of failed subjects in the first semester is also considered relevant. These significant attributes, along with the completion of complementary activities, may indicate that academic performance and engagement with the institution play a crucial role in preventing dropout. These attributes also clearly signal that newly enrolled students are more likely to be at risk of dropping out. Therefore, any institutional action should always consider this guideline.

The workload of the required complementary activities (CH_ATIV_COMPL_EXIGIDA) also plays a crucial role. Students who satisfactorily complete these activities seem to demonstrate a higher level of involvement and possibly a more solid commitment to the course. Moreover, it is worth considering the impact of socio-economic factors, such as monthly income and the number of family members, and to what extent they may influence the probability of dropping out. Low income or greater financial dependency may be risk factors. The age of the students can also be relevant, indicating different stages of maturity and engagement with studies. However, it was not possible to analyze whether older students reflect this observation and may have a lower risk of dropout, as they often face the need to balance work and studies.

The ratio between the completed workload and the total workload (RAZAO_CH_VENCIDA) can serve as an indicator of the student's academic progress. Those with a high proportion show consistent progress concerning the curriculum, which consequently can also contribute to mitigating the risk of dropout.

Figure 2. Main attributes according to the Random Forest algorithm.



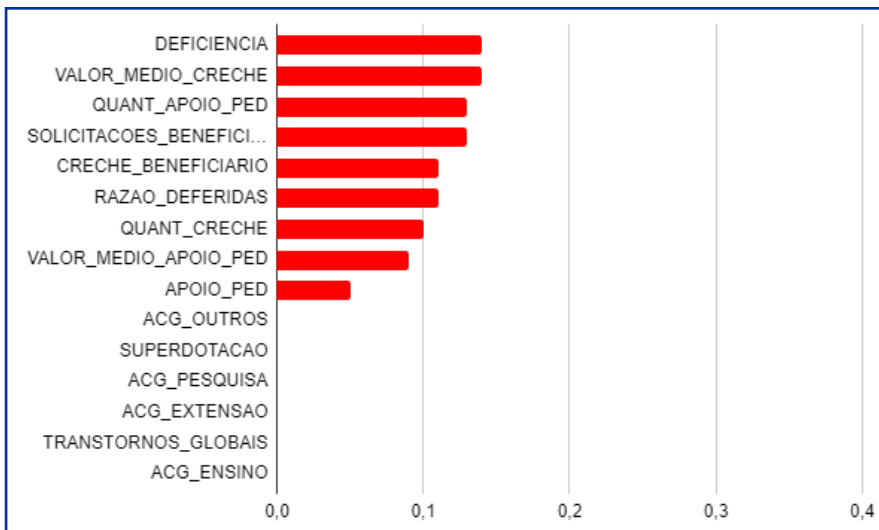
Source: Authors (2023).

In Figure 3, we can observe the attributes of lesser significance, i.e., those that scored below 0.14. The attributes associated with extracurricular activities (ACG_OUTROS, ACG_PESQUISA, ACG_EXTENSAO, ACG_ENSINO) did not receive considerable scores, which may primarily indicate limited data collection, possibly because students provide this information at the end of their course. Even students who eventually dropped out could have participated in such activities, but there may not have been records of these activities. Encouraging registration throughout the student's entire graduation process would likely increase the relevance of these attributes. On the other hand, attributes related to disabilities also showed low relevance. This may also reflect a reduced number of cases in the sampled population, contrasting with the previous scenario. In the former case, it is assumed that the problem is related to the timing of the records. As for students with disabilities, it is

plausible to speculate that the number of enrollments is small, which could make this attribute less relevant compared to others.

The remaining attributes in this segment are associated with student assistance, although this is a widely considered area when planning strategies to retain students in the academic environment. Benefits mainly related to pedagogical support and childcare services may not show a direct relationship with dropout, but they can serve as indicators of other socioeconomic factors. It would be necessary to investigate the number of aids granted to benefited students and evaluate their impact on the continuity of these students.

Figure 3. Less relevant attributes according to the Random Forest algorithm.



Source: Authors (2023).

Regarding the comprehensive analysis of the dataset, it was identified that 32 attributes in the sample exhibited a significance level below 0.2. Therefore, tests were conducted without the inclusion of these data in the samples to evaluate whether such exclusion had any impact on the overall performance of the algorithms. The results of these tests are documented in Table 2, as well as the tests conducted with 37 attributes (significance above 0.25) and 28 attributes (significance above 0.26). The tests were conducted using the Random Forest and C4.5 algorithms, and the

results were highly satisfactory, outperforming the tests conducted with the full sample. This resulted in an increase of more than 20% in accuracy for both algorithms.

It is observed that the Random Forest algorithm, which utilized a test base with 52 attributes, demonstrated the best performance in this experiment, achieving the highest accuracy rate for both dropout cases (“YES”) and non-dropout cases (“NO”) (highlighted row in Table 2).

It can be noted that as the number of attributes decreased in subsequent tests, there was a corresponding reduction in the effectiveness of the machine learning algorithms.

Table 2. Metrics of the algorithms tested with 52, 37, and 28 attributes.

Algorithm/Dataset	Metrics	Value	Confusion matrix (a= YES, b= NO)						
Random Forest with attributes above 0.2 (52 attributes)	Accuracy	84,4424%	<table border="1"><tr><td>a</td><td>b</td></tr><tr><td>4653</td><td>2281</td></tr><tr><td>1960</td><td>18366</td></tr></table>	a	b	4653	2281	1960	18366
	a	b							
	4653	2281							
	1960	18366							
Precision	0,842								
Recall	0,844								
F-Score	0,843								
C4.5 with attributes above 0.2 (52 attributes)	Accuracy	82,6082%	<table border="1"><tr><td>a</td><td>b</td></tr><tr><td>4528</td><td>2406</td></tr><tr><td>2335</td><td>17991</td></tr></table>	a	b	4528	2406	2335	17991
	a	b							
	4528	2406							
	2335	17991							
Precision	0,826								
Recall	0,826								
F-Score	0,826								
Random Forest with attributes above 0.25 (37 attributes)	Accuracy	84,4855%	<table border="1"><tr><td>a</td><td>b</td></tr><tr><td>4476</td><td>2145</td></tr><tr><td>1935</td><td>17742</td></tr></table>	a	b	4476	2145	1935	17742
	a	b							
	4476	2145							
	1935	17742							
Precision	0,843								
Recall	0,845								
F-Score	0,844								
C4.5 with attributes above 0.25 (37 attributes)	Accuracy	82,7059%	<table border="1"><tr><td>a</td><td>b</td></tr><tr><td>4410</td><td>2211</td></tr><tr><td>2337</td><td>17340</td></tr></table>	a	b	4410	2211	2337	17340
	a	b							
	4410	2211							
	2337	17340							
Precision	0,828								
Recall	0,827								
F-Score	0,828								
Random Forest with attributes above 0.26 (28 attributes)	Accuracy	82,9808%	<table border="1"><tr><td>a</td><td>b</td></tr><tr><td>4067</td><td>2286</td></tr><tr><td>2051</td><td>17079</td></tr></table>	a	b	4067	2286	2051	17079
	a	b							
	4067	2286							
	2051	17079							
Precision	0,828								
Recall	0,830								
F-Score	0,829								

Source: Authors (2023).

Considerations

Data mining is being used in higher education to analyze dropout rates and identify patterns that can help administrators or academic managers make more informed decisions. By analyzing data related to student information, data mining techniques can be used to identify students at risk of dropping out and plan appropriate interventions. Moreover, data mining can be used to improve evaluation processes and decision-making in higher education, using the knowledge generated as a basis for creating scenarios that characterize the university as a network of support and an enhancer for all students.

In this study, the proposed objective was achieved by clearly demonstrating that both the dataset and the selection of algorithms and data mining techniques play a crucial role in the accuracy of the results obtained. By analyzing the highlighted attributes, it is evident that seeking maximum information proliferation is not always the most advantageous strategy, as not all information is relevant in the knowledge discovery process. Additionally, attributes with limited data or a reduced number of instances lose their relevance in the context of machine learning. Therefore, one cannot generalize that the attributes considered less significant in this study would also be irrelevant in other evaluation scenarios.

A relevant conclusion that emerges is the importance of the university maintaining a solid and constantly updated database. It is imperative to create an organizational culture that promotes the quality of student records and the timely storage of information, as data about students' academic journeys will only be useful for studies that seek to identify dropout risks if recorded promptly. Recording elements such as curriculum components, extracurricular activities, or complementary courses only at the end of the academic journey makes conducting predictive dropout studies unfeasible.

Other data from the relationship between teachers and students can be used as attributes for this prediction, such as student attendance. This attribute is a common predictor in similar works, but for its effective use, it is essential to ensure that all teachers regularly record the presence and absence of students. Even if the omission occurs in only a small portion of teachers, it can cause an imbalance among students, impacting the algorithms' error.

By analyzing the effectiveness of the algorithms, it becomes evident that, despite the similarity of the metrics presented for the same dataset, there are variations in the accuracy of the algorithms. Therefore, the analysis of the confusion matrix becomes essential to guide the selection of the most appropriate algorithm. Algorithms that demonstrate higher accuracy and, most importantly, present higher values on the main diagonal should be chosen.

The Random Forest algorithm proved to be the best alternative among the solutions listed in this study, being consequently selected as the best choice for predictions aiming to identify students at risk of dropping out. This conclusion aligns with a wide range of studies that also identified this algorithm's high predictive potential. It remains for future researchers to explore parameterization strategies for the algorithm to analyze potential improvements in its performance. Strategies related to the database can also be analyzed, particularly concerning data preprocessing and missing data. The removal of attributes with many missing values and the standardization of scale, dimensions, or data nature can be considered.

References

- Alboaneen, D., Alenezi, M., Alrumayh, A., Almutairi, A., Alanazi, S., & Alotaibi, M. (2022). Development of a web-based prediction system for students' academic performance. *Data*, 7(2), 21. <https://doi.org/10.3390/data7020021>
- Al-Fairouz, E., & Al-Hagery, M. (2020). The most efficient classifiers for the student's academic dataset. *International Journal of Advanced Computer Science and Applications*, 11(9), 501–506. <https://doi.org/10.14569/IJACSA.2020.0110960>
- Alturki, S., Hulpuş, I., & Stuckenschmidt, H. (2020). Predicting academic outcomes: A survey from 2007 till 2018. *Technology, Knowledge and Learning*, 27(1), 275-307. <https://doi.org/10.1007/s10758-020-09476-0>
- Baker, R., De Carvalho, A., Da Costa, E., & Neves, P. (2011). Mineração de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, 19(2), 3-13. <https://doi.org/10.5753/rbie.2011.19.02.03>
- Borin, J. (2014). *Desenvolvimento de um software para análise de evasão na Unipampa Campus Bagé utilizando técnicas de mineração de dados* (Unpublished doctoral dissertation). Universidade Federal do Pampa, Bagé.
- Brandão, I. (2018). *Framework de mineração de dados educacionais em ambiente de cursos a distância governamental* (Unpublished undergraduate thesis). Universidade de Brasília, Brasília.

- Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: From concept to implementation*. Prentice-Hall, Inc.
- Castro, L. N., & Ferrari, D. G. (2016). *Introdução à mineração de dados: Conceitos básicos, algoritmos e aplicações*. Saraiva.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Advances in knowledge discovery and data mining. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *From data mining to knowledge discovery: An overview* (pp. 1-34). American Association for Artificial Intelligence.
- Fernandez-Garcia, A., Gonzalez, S., Rico, M., & Prieto, E. (2021). A real-life machine learning experience for predicting university dropout at different stages using academic data. *IEEE Access*, 9, 133076–133090. <https://doi.org/10.1109/ACCESS.2021.3119596>
- Flores, V., Heras, S., & Julian, V. (2022). Comparison of predictive models with balanced classes using the SMOTE method for the forecast of student dropout in higher education. *Electronics*, 11(3), 457. <https://doi.org/10.3390/electronics11030457>
- Franco, E., Martínez, R., & Domínguez, V. (2021). Predictive models of academic risk in computing careers with educational data mining. *Revista de Educación a Distancia*, 21(66). <https://doi.org/10.6018/red.425981>
- Hamoud, A., Hashim, A., & Awadh, W. (2018). Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(2), 26. <https://doi.org/10.9781/ijimai.2018.02.004>
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*. Elsevier/Morgan Kaufmann. <https://www.sciencedirect.com/science/book/9780123814791>
- Howlett, M., Ramesh, M., & Perl, A. (2013). *Política pública: Seus ciclos e subsistemas – uma abordagem integral*. Campus.
- Kabathova, J., & Drlík, M. (2021). Towards predicting student dropout in university courses using different machine learning techniques. *Applied Sciences*, 11(7), 3130. <https://doi.org/10.3390/app11073130>
- Lanot, A. (2012). *Mineração de dados aplicada na identificação da propensão à evasão na universidade* (Unpublished undergraduate thesis). Universidade Federal do Pampa, Bagé.
- Lee, S., & Chung, J. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences*, 9(15), 3093. <https://doi.org/10.3390/app9153093>
- Martins, B. (2017). *Uma discussão sobre diferentes ambientes de software para mineração de dados* (Unpublished undergraduate thesis). Universidade Federal do Maranhão, São Luís.
- Miranda, M., & Guzmán, J. (2017). Análisis de la deserción de estudiantes universitarios usando técnicas de minería de datos. *Formación Universitaria*, 10(3), 61-68. <https://doi.org/10.4067/s0718-50062017000300007>
- Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access*, 9, 140731–140746. <https://doi.org/10.1109/ACCESS.2021.3119596>
- Nascimento, F., de Carvalho, A. A., Lima, C. F., Neves, P. V., & Freitas, C. (2023). Mineração de dados educacionais: Uma revisão sistemática da literatura. *Zenodo*, 27(120), 1-21. <https://doi.org/10.5281/zenodo.7763620>

- Niyogisubizo, J., Nduwimana, A., Nzobonimpa, L., & Nkurunziza, D. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, 3, 100067. <https://doi.org/10.1016/j.caeai.2022.100067>
- Nuankaew, P., Namahoot, C., & Phewchewan, N. (2020). Prediction model of student achievement in business computer disciplines. *International Journal of Emerging Technologies in Learning (IJET)*, 15(20), 160. <https://doi.org/10.3991/ijet.v15i20.15273>
- Palacios, C., Arenas, M., Jimenez, E., & Villanueva, P. (2021). Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. *Entropy*, 23(4), 485. <https://doi.org/10.3390/e23040485>
- Pérez-Gutiérrez, B. (2020). Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico. *Revista UIS Ingenierías*, 19(1), 193-204. <https://doi.org/10.18273/revuin.v19n1-2020018>
- Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008). Data mining algorithms to classify students. In *Proceedings of the International Conference on Educational Data Mining* (pp. 8-17).
- SEMESP. (2023). Mapa do ensino superior 2023. <https://www.semesp.org.br/wp-content/uploads/2023/06/mapa-do-ensino-superior-no-brasil-2023.pdf>
- Siddique, A., Alam, S., & Marwah, A. (2021). Predicting academic performance using an efficient model based on fusion of classifiers. *Applied Sciences*, 11(24), 11845. <https://doi.org/10.3390/app112411845>
- Souza, V. F. (2021). Mineração de dados educacionais com aprendizagem de máquina. *Revista Educar Mais*, 5(4), 766-787. <https://doi.org/10.15536/reducarmais.5.2021.2417>
- Sunday, K., Jekayinoluwa, J., Adedokun, A., Ajao, T., & Yusuff, A. (2020). Analyzing student performance in programming education using classification techniques. *International Journal of Emerging Technologies in Learning (IJET)*, 15(2), 127-144. <https://doi.org/10.3991/ijet.v15i02.11527>
- Waikato, Departments of Computer Science and Software Engineering. (1999). *Weka 3: Machine learning software in Java*. <https://www.cs.waikato.ac.nz/ml/weka/>
- YAğCi, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1), 1-19. Springer Science and Business Media LLC. <http://dx.doi.org/10.1186/s40561-022-00192-z>