EvolveDTree: Analyzing Student Dropout in Universities

G.A.S.Santos[1], K.T.Belloze[1], L.Tarrataca[1], D.B.Haddad[1], A.L.Bordignon[2], D.N.Brandao[1]
[1]CEFET/RJ, Rio de Janeiro, RJ, Brasil
[2]UFF, Niterói, RJ, Brasil
gustavo.santos@eic.cefet-rj.br, {kele.belloze, tarrataca, diego.haddad, diego.brandao}@cefet-rj.br, alexb@id.uff.br

***Abstract*—Brazilian society suffers constant financial losses when higher education students disassociate from universities without completing the degree program in which they were enrolled. This is especially true when the institutions are funded through public resources. In order to minimize evasion losses, socioeconomic policies and programs were created to assist and support actions seeking to maximize the number of students that graduate in a suitable program time. This work presents a methodology that aims to predict evasion by using machine learning. Our approach was able to classify student abandonment with an average f-score and accuracy results above 95%. Our approach combines a decision tree alongside a genetic algorithm and cluster stratified sampling. The results obtained show that students with a Grade Point Average (GPA) below 5.79 and that have been enrolled for more than a year require careful monitoring because they tend to exceed the program duration time or abandon it. Furthermore, approximately one-third of all identified dropouts students occurred in the first year.**

***Keywords*—higher education, dropout, machine learning, evolutionary computation, features selection**

I. Introduction

Academic management of the Federal Institutes of Higher Education (FIHE) in Brazil undertakes various activities. Namely, these consist of teaching, research, and outreach programs. The latter of which is a concept aiming to promote interactions between communities and universities. Several academic management challenges exist which demand careful attention, *e.g.*, dropout, retention, unfilled vacancies [24], [27]

One of the main challenges of higher education is the dropout issue, which occurs when students abandon their studies without finishing the degree program into which they enrolled. This leads to significant financial losses in Brazilian society, a situation that is even more problematic when there is public financing of the institutions. In these cases, the vacancies assigned initially to these students become idle and need to be reallocated. Data from the 2016 census on Brazilian higher education shows that more than 10.6 million student places were offered in the degree program; of these: 73.8% were new spots, and 26.2% were absent vacancies [2]. Most institutes have actively tried to understand this phenomenon [26]. However, the number of wasted funds due to student dropout are at unacceptable levels [5].

According to an OECD (Organisation for Economic Co-operation and Development) report [20], the average annual cost per undergraduate enrolled in the Brazilian higher education system was US$ 13,539.90. The 2017 census on higher education [6] also showed the magnitude of the problem, with the number of dropped students reaching 1,818,838.

978-1-7281-7539-3/20/$31.00 ©2020 IEEE

Given the financial and social impact of student dropout, this work presents an approach that aims to tackle the issue by using machine learning based on decision trees with genetic algorithms. Our model can be used to predict whether a student will evade or not. Our main objectives are to provide insights into potential students who may dropout to (i) give improvements to academic management of undergraduate programs and (ii) prevent at-risk students from evasion.

This paper is organized as follows: Section II presents the primary bibliographical references within the context of the problem being analyzed; Section III presents a decision tree model based on genetic algorithms, capable of classifying dropout profiles which aims to identify at-risk students preventively; Section IV presents the results obtained; Section V presents the final considerations alongside future work proposals.

II. Related Works

Dropout, and the associated root causes, are issues that have been thoroughly researched in teaching institutions in Brazil and elsewhere [26]. The research fields responsible for tackling the dropout issue, alongside the respective leading causes, can be found in an Educational Data Mining (EDM) systematic review published in [24]. For instance, Andriola [1] draws attention to the alarming rates of degree program transfers within Brazilian institutions. This points to weaknesses in vocation identification procedures. Besides, it also represents a burden to society, since unduly occupied student vacancies often translate into inefficient investments. Another critical issue is the dropout rate. As mentioned in [21], dropout rates are often underestimated when institutions are performing internal degree program performance evaluations. Also, the author states that 64% of students who abandoned their programs eventually obtained an equivalent degree in another institute. One reason for this is that student mobility is allowed within FIHEs.

In [8], students, and managers from FIHEs belonging to several regions of Brazil were interviewed. The author mentions inconsistencies between the official evasion numbers and the ones presented by the managers during the interviews. According to the author, several of the interviewees were unable to discuss evasion numbers since this topic is perceived as taboo.

In 2012, UNESCO published a work focusing on the challenges and perspectives for Brazilian higher education [27]. This study emphasized the need for researchers studying the specificities of the Brazilian issues to focus more attention on idle student places and evasion. The same work also mentions that the main reasons for unoccupied spots and dropout remain widely misunderstood, namely: (i) insufficient financial resources, (ii) recent diversification and (iii) system quality.

The work of Manhaes *et al.* [16] aims to assist in institutional management. Data from undergraduate students of a significant educational institute were analyzed using the following classification algorithms: Decision Trees, Support Vector Machines, Naive Bayes, and Multi-layer Perceptron Neural Network. The Naive Bayes model was used to present a quantitative approach. According to the authors, the use of EDM methods helps to identify the students still enrolled with a higher probability of dropout. Also, they offer the institute an additional analysis of the high dropout rate problem. Maschio's research [17] presents an up to date panorama of the studies being developed in Brazil regarding EDM.

Our work differentiates from the previously mentioned ones for the following reasons: (i) innovative feature selection process, (ii) overfitting minimization, (iii) construction of decision tree sets with specific rules concerning knowledge area or undergraduate programs, and (iv) differential preprocessing based on cluster sampling approach.

III. The EvolveDTree Model

EvolveDTree stands for "Evolutionary Decision Tree", a model that was developed using techniques and strategies commonly employed in Machine Learning (ML). ML attempts to develop computational-based learning processes [12]. Among several ML techniques, Decision Trees (DT) have proven to be a valuable technique in describing, classifying, and generalizing data and have been used in diverse areas [18], [25].

Another ML technique are Genetic Algorithms (GA), which are search and optimization methods inspired by the evolutionary mechanisms of living beings [9], [11]. According to [14], these algorithms can be used to improve the induction procedure of decision trees, namely, in what concerns: (i) search method for feature selection and (ii) data partitioning process into sub-trees. This work focuses on reducing data unbalancing and improving accuracy to achieve a better classification result. This section presents the details of the model developed, which is capable of mining educational data to predict evasion by combining DT and GA.

The following sections are organized as follows: Section III-A describes the data set employed, Section III-B presents the evasion behaviors that were identified by graphical analysis; Section III-C discusses data preprocessing; Section III-D elaborates on feature selection and classification;

A. Data Set

The dataset applied in this research contains student information as follows: university admittance exam grade, academic record, public policy record, race-ethnicity, and sociodemographic data. The dataset includes 12969 instances and 28 features of students from the 106 undergraduate face-to-face courses that are offered in a larger Brazilian FIHE. The complete description of this data is available in https://github.com/gassantos/evolvedtree/datasetinfo. This dataset comprehends students ingresses between the years of 2012 until 2014, and those that dropped out or graduated up to 2018. More details about the creation of this dataset can be seen in [23].

B. Exploratory Data Analysis

This section describes the insights that were gained by the plots and the tables produced in the quantitative data analysis. During this stage, some important characteristics were observed. Table I presents the correlation matrix, that describes the dataset influence percentages regarding the "**GPA**" and "**FinalStatus**" features. The results are sorted from highest to lowest value.

TABLE I
Dataset Correlation Matrix for the GPA and FinalStatus features.

Feature	GPA	FinalStatus
GPA	1.000000	0.6337970
TotalCourseHrs	0.702870	0.9017570
FinalStatus	0.633797	1.0000000
YearsEnrolled	0.539021	0.4808890
ExamLang	0.141993	0.0992620
ExamWrite	0.139197	0.0956480
ExamScien	0.103915	0.0408970
ExamHuman	0.094311	0.0183409
ExamMath	0.067701	0.0362350
IdShift	0.076495	−0.014530
Withdraws	0.037985	−0.008887
IdShiftCurr	0.008767	−0.057659
ApplicYear	−0.09314	−0.208442
ApplicSemester	−0.10859	−0.101707
Age	−0.14578	−0.067144

The second observation was that the dataset had an unbalanced class representation, with 76% of students having dropped out and 24% having graduated. This unbalanced representation tends to hinder the learning process of the model. Moreover, Table II presents details quantifying the amount of data and exhibits for each class the median values for the features: 'Age' (**Age**), 'ClosureSemester' (**Cs**), 'GradePointAverage' (**GPA**), 'TotalAmountCourseHours' (**Tach**) and 'YearsEnrolled' (**Nye**). Based on the results, it is possible to see that the biggest dropout values occur in the first semester of each school year by students with age close to 25 years, and with median values for GPA and the amount hours course of, respectively, 3.4 and close to 240 curriculum hours.

Another observation is derived from the histogram shown in Figure 1. Namely, Brazilian FIHE can automatically cancel student enrollment whose academic performance is low. Consequently, the samples presented in the GPA regions between 0 and 3.0 show that approximately 34% of students are probably dropped in the first year of enrollment. This pattern is related

TABLE II
DATASET CHARACTERIZATION ACCORDING TO CARDINALITY AND MEDIAN VALUE FOR 'AGE', 'CLOSURESEMESTER' (CS), 'GRADEPOINTAVERAGE' (GPA), 'TOTALAMOUNTCOURSEHOURS' (TACH) AND 'YEARSENROLLED' (NYE).

Instance	Count	Age	Cs	GPA	Tach	Nye
Dropped	9852	25	1º	3.4	240	3
Graduated	3117	24	2º	8.3	3199	5

to automatic cancellation rules, which are based on poor performance on courses, direcly translating to a low GPA.

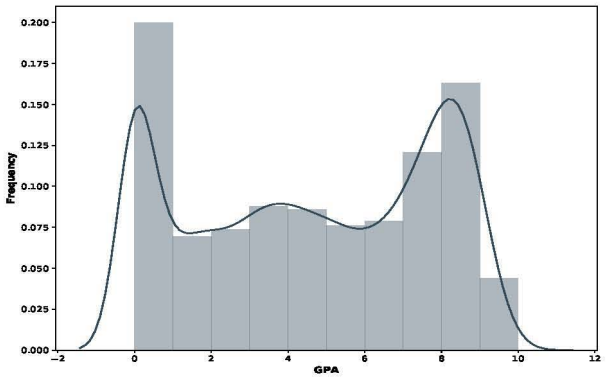

Fig. 1. The histogram is representing the sample of all students', including graduated and dropped, through the ranges of GPA values and its frequency students per each one GPA.

After that, another observation is related to the student distributions and their ethnicity. Figure 2 presents a boxeplot relating GPA and ethnicity. This plot illustrates that for most ethnicities, the median GPA value is between 4 and 6. The exception to this rule is the yellow-labeled ethnicity. However, this ethnicity is only represented by 31 instances.

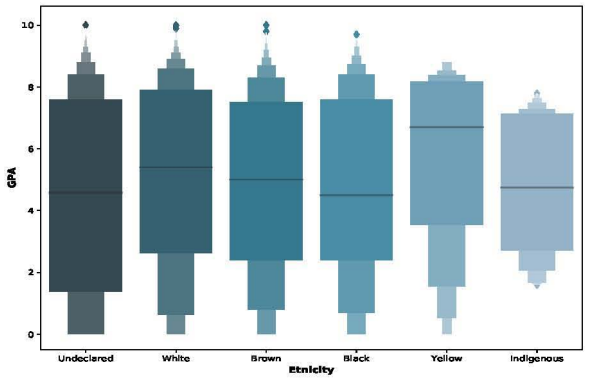

Fig. 2. Boxen plot graphs representing specifically each of the ethnicity belonging to the students and relating them to the GPA performance referring for the respective ethnicity.

C. Data preprocessing

The preprocessing applied in this work is based on unsupervised learning to stratify the data set [4]. This stratification process is based on clusterization through the *K-means* algorithm, which aims to partition the dataset into 10 clusters. This value was chosen because we prefer to have a good number of samples for the training stage. Nine clusters were used for the training stage with 90% of data, thus favoring a better learning scenario. The last cluster was used for testing.

D. Feature Selection

This stage performs feature selection (FS) by employing *Evolutionary Algorithms* [10], through the DEAP Framework [3], whose aim is to identify the feature set that is most representative for a student. Techniques exist that can be used to identify and remove unnecessary features that fail to contribute, or harm, the performance of the predictive model [7], [15]. Several works have shown that the following combination of features is valid for improving performance: (i) student record; (ii) socioeconomic information; (iii) family data; as well as (iv) academic performance [28].

The GA developed for this work implemented an heuristic to perform FS responsible for optimizing data set size and discarding less relevant features. This strategy is similar to the one described in [28], which is based on the process of improving fitness. In this technique, at a given iteration, the selected classification algorithm is trained on n input features. During this process, the heuristic discards the features that have the lowest performance and evaluates the remaining set of features to get an accuracy greater than or equal to the previous step.

In each generation, the best individual is selected based on maximum accuracy. The feature subset of this individual is then provided as input for the feature selection GA. New individuals are then generated based on these, which allows the algorithm to converge. The GA results can be seen in the section IV. This FS allows for a reduction in the number of dimensions which minimizes overfitting since we applied DT for dropout classification. The DT implementation used in this work was presented in [22].

IV. RESULTS

This section details the results obtained by each stage of our model. Accordingly, the numerical results described show the improvements obtained for: (i) the preprocessing stage (Section IV-A); (ii) the training and FS stage (Section IV-B; and (iii) evaluation of the EvolveDTree model (Section IV-C).

A. Preprocessing Process

The holdout technique was employed with a partitioning ratio of respectively 80% and 20% to create training and test sets in the initial experiments. In these early experiments, an average accuracy of 0.73 and an average Matthews correlation coefficient (MCC) of 0.17 were achieved. Based on the values, we opted to improve the preprocessing stage. This was done since the dropped students' class was overly represented in

the data set. Namely, the data set contained 76% of dropped students and 24% of graduated students. As a result, we chose to use a data clusterization technique and then apply cross-validation to the 10 clusters. This technique allowed for an increase in accuracy of approximately 17% while the average MCC increased to 45%.

B. Evaluation of Feature Selection

By applying the GA we are able to obtain a feature subset with better accuracy, when compared against the baseline value. Table III presents the results from the execution of the GA with each column representing, respectively: "Generation number" (**Gen.**), "Number of individuals" (**Ind.**), "Average accuracy" (**Avg.**), "Standard deviation accuracy" (**Std.**), "Minimum accuracy" (**Min.**) and "Maximum accuracy" (**Max.**).

TABLE III
GENERATION EVALUATION.

Gen.	Ind.	Avg.	Std.	Min.	Max.
0	100	0.914149	0.0943826	0.680964	0.991807
1	53	0.975846	0.0427739	0.751807	0.992289
2	56	0.986593	0.0254018	0.766265	0.992289
3	61	0.986824	0.0259713	0.762892	0.992771
4	50	0.988241	0.0195519	0.810120	0.992771
5	54	0.990318	0.0069133	0.922410	0.992771
6	47	0.990525	0.0077981	0.913735	0.992771
7	68	0.991089	0.0017045	0.985542	0.993253
8	62	0.989320	0.0202991	0.787952	0.993253
9	73	0.988448	0.0319965	0.670361	0.993253
10	51	0.992048	0.0012201	0.986506	0.993253

The results produced by the FS process promoted significant gains. In the training stage, a baseline best accuracy of 0.993253 was obtained. During the individual's generation process in the validation stage, the best accuracy was 0.99460. This value was obtained with an individual using the following feature set:

- SocProgram;
- ClosureSemester;
- ExamWrite;
- ApplicYear;
- GPA;
- Ethnicity;
- ApplicSemester;
- NumYearsEnrolled.

The evolutionary heuristic discarded 20 features from a set of dimensionality 28. Figure 3 presents the accuracy obtained for the validation and test sets used in the FS process. The minimum and optimal values achieved were, respectively, 0.680964 and 0.993253.

Besides evaluating the fitness function of the GA, we also tested DT classifiers, logistical regression, and K-Nearest Neighbours (KNN). The evaluation metrics used were precision, F-Score, and accuracy. The best result obtained for the fitness function was achieved through DTs.

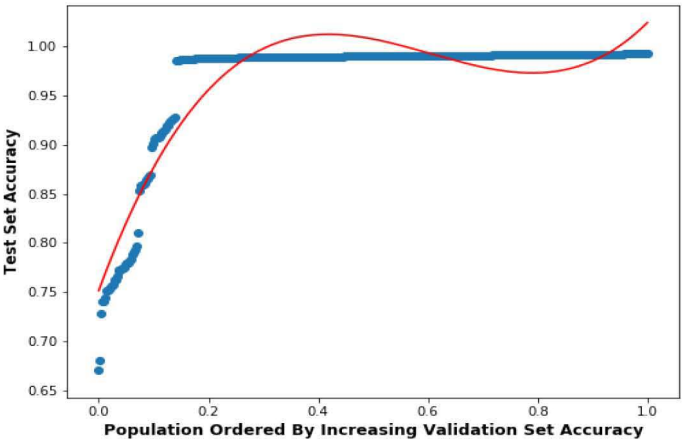

Fig. 3. The red curve represents the performance interpolation of GA in test data. The blue curve represents validation data of the FS process.

C. Evaluation of EvolveDTree Model

The DT obtained by our *EvolveDTree* model is presented in Figure 4. Model evaluation is performed through similar techniques that are detailed in [13]. We chose to focus on performance evaluation metrics that are highlighted in [19].

Given the attributes presented by the decision nodes, it is possible to conclude that students with a GPA below 5.79 are highly likely to evade. Other observations are also pertinent, *e.g.:* the root node partitions the data into two subtrees, in accordance with the number of samples. The results produced in this evaluation are presented separately in Tables IV and V. These discriminate for each algorithm the values of the metrics that were used, namely, '**F-Score**', '**Precision**', '**Accuracy**', '**Mcc**', '**Kappa**' and '**R-auc**'.

TABLE IV
METRICS EVALUATION FOR EVOLVEDTREE AND OTHER ALGORITHMS REGARDING DROPOUT CLASSIFICATION.

Algorithms	F-Score	Precision	Accuracy
EvolveDTree	0.98978	0.98225	0.99383
KNeighbors	0.98530	0.97842	0.99290
AdaBoost	0.98526	0.98086	0.99290
SVC	0.98207	0.97707	0.99136
MLP	0.97283	0.95533	0.98674
RandomForest	0.96926	0.98537	0.98550
QDA	0.92435	0.86031	0.96083
NaiveBayes	0.91573	0.84456	0.95590

Table VI presents the set of results produced by *EvolveDTree* for the two classification classes. The table discriminates for each of one of the classification classes, respectively, '**Dropped**' and '**Graduated**', the numerical values obtained through the following metrics: "Precision" (**Prec.**), "Recall" (**Rec.**), "F-Score" (**F**), "Support" (**Sup.**), "True Positive" (**TP**) and "False Positive" (**FP**). By considering the precision obtained for each of these we are able to obtain an average precision of 0.96 for our *EvolveDTree* method.

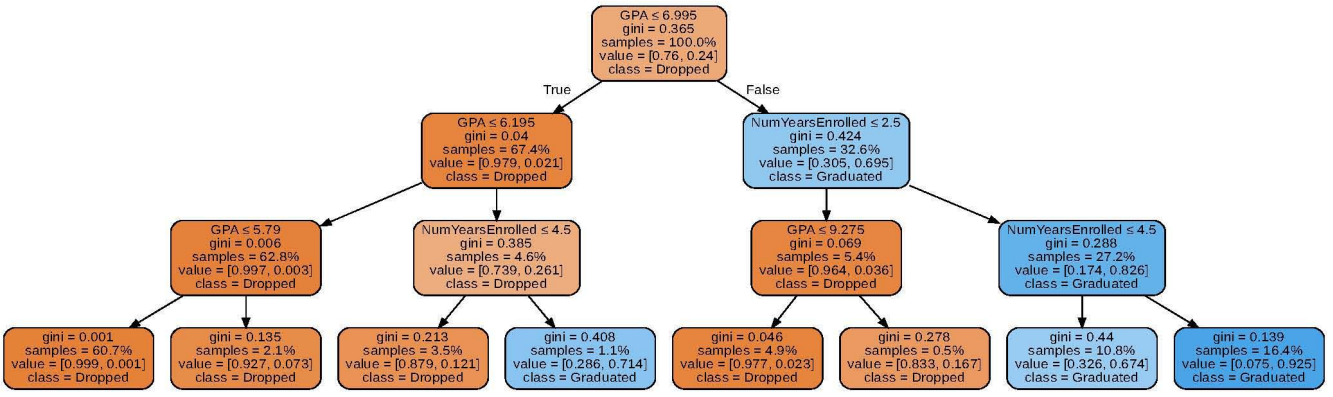

Fig. 4. The Decision Tree produced by *EvolveDTree*.

TABLE V
METRICS EVALUATION FOR EVOLVEDTREE AND OTHER ALGORITHMS REGARDING DROPOUT CLASSIFICATION.

Algorithms	MCC	Kappa	R-auc
EvolveDTree	0.98658	0.98653	0.99587
KNeighbors	0.98067	0.98062	0.99269
AdaBoost	0.98061	0.98059	0.99181
SVC	0.97641	0.97638	0.98991
MLP	0.96435	0.96407	0.98819
RandomForest	0.96000	0.95978	0.97460
QDA	0.90268	0.89813	0.97380
NaiveBayes	0.89195	0.88615	0.97100

TABLE VI
CLASSIFICATION REPORT FOR *EvolveDTree*

Class	Prec.	Rec.	F	Sup.	TP	FP
Dropped	0.97	0.98	0.98	985	966	19
Graduated	0.94	0.92	0.93	311	273	38

V. FINAL REMARKS

This paper presented an EDM model capable of classifying evasion, focusing on Brazilian FIHE data. Our work presented a machine learning methodology combining decision trees, genetic algorithms, and cluster stratified sampling in the same approach.

EvolveDTree showed efficient performance and good accuracy. This is important since it allows for an education institute to minimize the expenses associated with the evasion problem, given that at-risk students can be adequately targeted. *EvolveDTree* is a solution that can be used by a FIHE, with minor adjustments required to ensureque data adherence. The implementation project [1] of this EDM work is available online for all interested, under the condition of referencing it when used.

As a result, the *EvolveDTree* model displayed better performance when compared against classical ML methods found in the literature. The work considers some specific attributes: ethnicity (race/skin color), social program, and university admittance-exams. The classification model had a gain of 13% due to the use of an evolution-based FS method. The preprocessing stage also gave a significant model improvement in the imbalanced data treatment with up 45% of the MCC measure gain due to the use of the sample balancing method proposed based on unsupervised learning.

Also, this approach was able to produce a better choice of decision tree rules and also reduced imbalanced data. An exploratory analysis of the dataset was made. Feature selection was performed using genetic algorithms aiming to remove any features that would harm the performance of the predictive model.

To Future works, the methodology proposed will be applied to other scenarios from diverse FIHE, to confirm its generality. After that, new research directions could try to ask some questions. First, how is estimated the quality of education and its connection to students' dropout? Second, what measures might be taken to decrease student's dropout?

[1]Source code – https://github.com/gassantos/evolvedtree

ACKNOWLEDGMENT

The authors are grateful for the support of CNPq - National Council for Scientific and Technological Development, FAPERJ - Foundation for Research Support of the State of Rio de Janeiro and CAPES - Coordination for the Improvement of Higher Education Personnel for the funding of this research and the Universidade Federal Fluminense (UFF) for providing the data through its organizational units: Superintendency of Information Technology (STI) and Pro-Rectory of Undergraduate.

REFERENCES

[1] W. B. Andriola. Evasão discente na universidade federal do ceará (ufc): proposta para identificar suas causas e implantar um serviço de orientação e informação (soi). *Ensaio. Avaliação de Políticas Públicas em Educação, Rio de Janeiro,(11)*, 40:332–347, 2003.

[2] D. de Estatísticas Educacionais (DEED). Censo da educação superior - notas estísticas 2016. *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)*, 2016. Acessado em 13 de Outubro de 2018.

[3] F.-M. De Rainville, F.-A. Fortin, M.-A. Gardner, M. Parizeau, and C. Gagné. Deap: A python framework for evolutionary algorithms. In *Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '12, pages 85–92, New York, NY, USA, 2012. ACM.

[4] N. Diamantidis, D. Karlis, and E. A. Giakoumakis. Unsupervised stratification of cross-validation for accuracy estimation. *Artificial Intelligence*, 116(1-2):1–16, 2000.

[5] C. A. dos Santos Baggi and D. A. Lopes. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior*, 16(2), 2011.

[6] U. F. Fluminense. Censo 2017 - evasão. Available at http://www.uff.br/?q=censo-2017-evasao. Accessed: 2019-02-22.

[7] H. Frohlich, O. Chapelle, and B. Scholkopf. Feature selection for support vector machines by means of genetic algorithm. In *Tools with artificial intelligence, 2003. proceedings. 15th ieee international conference on*, pages 142–148. IEEE, 2003.

[8] N. d. L. Gaioso. O fenômeno da evasão escolar na educação superior no brasil. *Brasília, DF: Universidade Católica de Brasília*, 2005.

[9] D. E. Goldberg and J. H. Holland. Genetic algorithms and machine learning. *Machine learning*, 3(2):95–99, 1988.

[10] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[11] J. H. Holland. Genetic algorithms. *Scientific american*, 267(1):66–73, 1992.

[12] N. Japkowicz and M. Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

[13] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[14] J. R. Koza. Concept formation and decision tree induction using the genetic programming paradigm. In *International Conference on Parallel Problem Solving from Nature*, pages 124–128. Springer, 1990.

[15] R. Leardi. Application of genetic algorithm–pls for feature selection in spectral data sets. *Journal of Chemometrics*, 14(5-6):643–655, 2000.

[16] L. M. B. Manhães. Predição do desempenho acadêmico de graduandos utilizando mineração de dados educacionais. *Doutorado em Engenharia*

[25] H. Sharma and S. Kumar. A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, 5(4):2094–2097, 2016.

de Sistemas e Computação Instituição de Ensino: Universidade Federal do Rio de Janeiro, Rio de Janeiro. Biblioteca Depositária: BIBLIOTECA DO CT, 2015.

[17] P. Maschio, M. A. Vieira, N. Costa, S. de Melo, and C. P. Júnior. Um panorama acerca da mineração de dados educacionais no brasil. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 1936, 2018.

[18] S. K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data mining and knowledge discovery*, 2(4):345–389, 1998.

[19] F. A. Narudin, A. Feizollah, N. B. Anuar, and A. Gani. Evaluation of machine learning classifiers for mobile malware detection. *Soft Computing*, 20(1):343–357, 2016.

[20] P. A. M. M. Nascimento and R. E. Verhine. Considerações sobre o investimento público em educação superior no brasil. *Instituto de Pesquisa Econômica Aplicada (IPEA)*, 2017.

[21] A. S. Paredes. *A evasão do terceiro grau em Curitiba*. NUPES, 1994.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[23] G. A. S. Santos, A. Bordignon, D. Haddad, D. Brandão, L. Tarrataca, and K. Belloze. Data warehouse educacional: Uma visão sobre a evasão no ensino superior. In *Anais do XXXIV Simpósio Brasileiro de Banco de Dados*, pages 235–240, Porto Alegre, RS, Brasil, 2019. SBC.

[24] G. A. S. Santos, A. L. Bordignon, S. L. G. Oliveira, D. B. Haddad, D. N. Brandão, and K. T. Belloze. A brief review about educational data mining applied to predict student's dropout. In *Anais da V Escola Regional de Sistemas de Informação do Rio de Janeiro*, pages 86–91, Porto Alegre, RS, Brasil, 2018. SBC.

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

[27] P. Speller, F. Robl, and S. M. Meneghel. Desafios e perspectivas da educação superior brasileira para próxima década. *Oficina de Trabalho. p. 164, 2012. ISBN: 978-85-7652-171-6*, 2012.

[28] F. Tan, X. Fu, Y. Zhang, and A. G. Bourgeois. A genetic algorithm-based method for feature subset selection. *Soft Computing*, 12(2):111–120, 2008.