A Genetic Algorithm with Flexible Fitness Function for Feature Selection in Educational Data: Comparative Evaluation

Danielle Albuquerque, Luís Tarrataca, Diego Brandão, Rafaelli Coutinho

Federal Center of Technological Education of Rio de Janeiro - CEFET/RJ
danielle.albuquerque@aluno.cefet-rj.br
{luis.tarrataca, diego.brandao, rafaelli.coutinho}@cefet-rj.br

Abstract.

Educational Data Mining is an interdisciplinary field that helps understand educational phenomena through computational techniques. The databases of educational institutions are usually extensive, possessing many descriptive attributes that make the prediction process complex. In addition, the data can be sparse, redundant, irrelevant, and noisy, which can degrade the predictive quality of the models and affect computational performance. One way to simplify the problem is to identify the least important attributes and omit them from the modeling process. This can be performed by employing attribute selection techniques. This work evaluates different feature selection techniques applied to open educational data and paired alongside a genetic algorithm with a flexible fitness function. The methods and results described herein extend a previously published paper by: *(i)* describing a larger set of computational experiments; *(ii)* performing a hypothesis test over different classifiers; and *(iii)* presenting a more in-depth literature revision. The results obtained indicate an improvement in the classification process.

Categories and Subject Descriptors: H.2 [**Computing methodologies**]: Miscellaneous; H.3 [**Machine learning algorithms**]: Miscellaneous; I.7 [**Feature selection**]: Miscellaneous

Keywords: Feature Selection, Genetic Algorithm, Educational Data Mining

1. INTRODUCTION

The advent of the Internet and the digitalization of physical records resulted in a new reality for the educational field. Some examples of the technologies that led to the creation of large data repositories are: *(i)* educational software; *(ii)* public databases; *(iii)* computerized school management systems; and *(iv)* remote lectured courses. The area of Educational Data Mining (EDM) has emerged from this context as a way to assist in understanding and analyzing such data. One of the main objectives of EDM is to provide tools that: *(i)* enable mapping of student profile; *(ii)* analyze evasion risk; and *(iii)* determine the main factors impacting academic performance. EDM enables more efficient financial and pedagogical investments [Romero and Ventura 2010]. As in other domains, educational data can be sparse, redundant, irrelevant, and noisy. These factors can hamper the quality and computational performance of exploratory and predictive models [Farissi et al. 2020]. One possible way to minimize these issues is to identify less important features, using Feature Selection (FS) techniques, and remove them from the modeling process.

A diverse set of FS techniques has been employed in the education field. Namely, due in part to its simplicity, many works employ filtering methods with statistical ranking criteria [Ramaswami and Bhaskaran 2009; Rachburee and Punlumjeak 2015; Sasi Regha and Uma Rani 2015; Abid et al.

2018; Gopalakrishnan et al. 2018; Febro 2019; Huang et al. 2019; Dimic et al. 2019; Enaro and Chakraborty 2020; Das et al. 2020; Chaudhury and Tripathy 2020; Muchuchuti et al. 2020; Chaves et al. 2021]. Other methods employ wrapper approaches that use genetic algorithms (GA) to select the best set of attributes [Wafi et al. 2019; Almasri et al. 2020; Farissi et al. 2020; Santos et al. 2020]. Other works include FS as part of a classification model, and are commonly referred to as embedded approaches [Gitinabard et al. 2018; Niu et al. 2018; Hassan et al. 2019; Teodoro and Kappel 2020]. Furthermore, comparisons between different FS techniques have been examined in order to identify the most suitable to specific scenarios [Punlumjeak and Rachburee 2015; Zaffar et al. 2018; Hashemi et al. 2018; Ajibade et al. 2019; Ahmed et al. 2019; Govindasamy and Velmurugan 2019; Ahmed et al. 2020; Chaves et al. 2021; Jalota and Agrawal 2021].

The aforementioned work only present classical FS techniques without proposing an approach specific to educational data. This work presents an evaluation of different filter and classifiers techniques, combining them with the GA proposed in [de Albuquerque et al. 2021]. The set of results published in the latter are extended herein with a more comprehensive set of experimental evaluations. In addition, a hypothesis test is also performed in order to determine if the differences in results, obtained using distinct classifiers (in conjunction with the GA), are statistically significant. A more detailed review of the state of the art is also presented. The methodology proposed allows educational experts to have greater flexibility in inserting practical information into the FS process. This is performed in accordance with the needs and realities that are characteristic of the educational environment. This approach allows for the intrinsic characteristics of each school environment to be better explored.

This paper is organized as follows: Section 2 presents a brief FS review. Section 3 describes the literature review about FS techniques in an EDM context. Section 4 presents our novel version that combines FS (Chi-Square and Correlation) with a GA and some classifiers (K-Nearest Neighbors, Decision Tree, Naive Bayes, Logistic Regression, Random Forest and Neural Networks). Section 5 covers the experimental analysis performed. Section 6 presents the conclusions and describes future work.

2. THEORETICAL BACKGROUND

Dimensionality reduction (DR) is a preprocessing stage that attempts to find redundant and irrelevant attributes of a database with the purpose of reducing their number [Han et al. 2012]. DR can be performed in two ways, namely, through feature extraction (FE) and feature selection (FS). The classification of the different DR methods is presented in Figure 1. FE combines the attributes through linear and non-linear transformations as a means to create a new set of attributes, with a smaller cardinality than the original one, but that is still able to convey the same information. FS chooses an attribute subset of the original dataset with the objective of minimizing redundancy and maximizing attribute relevance in regards to a target attribute. FS can be performed through three approaches, namely [Han et al. 2012; Ullah et al. 2017; Velliangiri et al. 2019; Venkatesh and Anuradha 2019]: filter, wrapper and embedded.

The filter approach uses statistical metrics from the data in order to rank attributes and select the k best performing ones. The wrapper method selects features through an optimization algorithm that creates attributes sets and selects the one that is best evaluated by an objective function (OF). The embedded approach gathers the qualities of filter and wrapper methods, enabling FS to be embedded in the classification (or regression) process [Dash and Liu 1997; Kohavi and John 1997; Guyon and Elisseeff 2003; Wang et al. 2015].

FS allows for several advantages, such as: *(i)* improve algorithmic efficiency; *(ii)* enhance classification precision; *(iii)* improve data quality; *(iv)* avoid overfitting; and *(v)* facilitate result visualization [Chandrashekar and Sahin 2014]. FS is a NP-complex problem [Davies and Russell 1994; Chen et al. 1997], with multiple techniques having been proposed, such as the GA wrapper approach and the filtering methods Chi-Square and Pearson correlation [Tan et al. 2008; Febro 2019], in which they

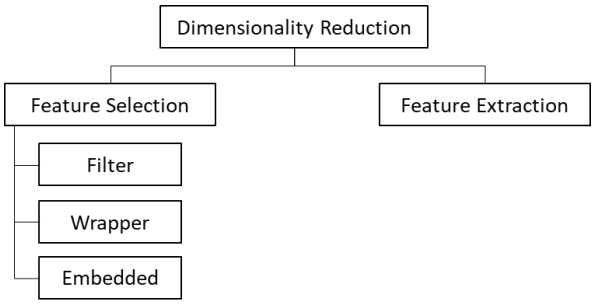

Fig. 1. Classification of dimensionality reduction methods (Source: [Chetana et al. 2020]).

are employed in the methodology of this work and are detailed below.

Genetic Algorithm (GA) is a meta-heuristic based on the principle of evolution. The candidate solutions, referred to as individuals, are evaluated by an Objective Function (OF) responsible for calculating their respective fitness values. The idea is for the individual with the highest fitness to "survive" and its information carried over to the next generation. During each generation, the fittest individuals are crossed-over and allowed to suffer mutations with a certain probability. This process results in a new population of individuals that can be used in the next generation. This procedure is performed until a halting criterion is triggered (*e.g.*, a certain number of generations has been reached or a predetermined objective function value was obtained)[Tan et al. 2008; Santos et al. 2020].

The filtering methods tend to be simpler and faster. They are guided by a statistical evaluation metric that ranks the attributes in decreasing order of importance. Among these methods, the Chi-Square (CS) test for FS is a technique frequently employed in educational data [Febro 2019]. The CS test for FS is calculated through the equation $CS = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$ and uses of the observed and estimated frequencies between attribute and class, where n is the quantity of possible categories for the attribute, k is the quantity of classes, A_{ij} is the observed frequency of category i in class j and E_{ij} is the estimated frequency, calculated by the data distribution amongst attribute and class.

Pearson correlation is another statistical method used in FS to measure the relationship of each input attribute with the target attribute (class). This measure is calculated according to the equation $P(X, Y) = \frac{Cov(x,y)}{\sigma_X \sigma_Y}$, where X is the input attribute and Y is the class. The equation uses the covariance between them ($Cov(x, y)$) and the standard deviation of each of them σ_X and σ_Y [Prabha et al. 2019].

3. RELATED WORK

The literature review that is to follow consists of the relevant works concerning the use of FS techniques for EDM research. The set of references mentioned was obtained by searching the Scopus database in January 2022 using the query ("*Educational data mining*" AND ("*Feature selection*" OR "*Dimensionality reduction*")). The search considered the title, abstract, and keywords and was limited to articles in the English language. A total of 120 documents were returned, of which 33 articles were selected because they had FS in educational data as their main objective. This set of works utilized a diverse set datasets, namely: student questionnaires [Ahmed et al. 2019], data stemming from e-learning platforms [Govindasamy and Velmurugan 2019; Jalota and Agrawal 2021], school databases [Abid et al. 2018; Sokkhey and Okazaki 2020] and university graduation departments [Ahmed et al. 2020; Chaves et al. 2021]. The selected articles can be grouped in accordance with the FS method employed, namely: *(i)* filtering methods; *(ii)* wrapper approaches; *(iii)* embedded approaches; *(iv)* methods for feature extraction; and *(v)* FS comparison studies, as summarized in Table I.

Filtering methods with statistical ranking criteria are the most widely used due to their simplicity

Table I. Related works summarized by method used.

Work	Feature Selection			Feature Extraction	Comparison Studies
	Filter	Wrapper	Embedded		
[Ramaswami and Bhaskaran 2009]	X				
[Punlumjeak and Rachburee 2015]	X	X	X		X
[Rachburee and Punlumjeak 2015]	X				
[Sasi Regha and Uma Rani 2015]	X				
[Abid et al. 2018]	X				X
[Gitinabard et al. 2018]			X		
[Gopalakrishnan et al. 2018]	X				
[Hashemi et al. 2018]	X	X			X
[Niu et al. 2018]			X		
[Zaffar et al. 2018]	X			X	X
[Ahmed et al. 2019]	X	X			X
[Ajibade et al. 2019]	X	X			X
[Dimic et al. 2019]	X				
[Febro 2019]	X				
[Govindasamy and Velmurugan 2019]	X			X	X
[Hassan et al. 2019]			X		
[Huang et al. 2019]	X				
[Wafi et al. 2019]		X			
[Ahmed et al. 2020]	X	X			X
[Almasri et al. 2020]		X			
[Chaudhury and Tripathy 2020]	X				
[Das et al. 2020]	X				
[Enaro and Chakraborty 2020]	X				
[Farissi et al. 2020]		X			
[Muchuchuti et al. 2020]	X				
[Poudyal et al. 2020]				X	
[Santos et al. 2020]		X			
[Šarić Grgić et al. 2020]				X	
[Sokkhey and Okazaki 2020]	X				
[Teodoro and Kappel 2020]			X		
[Chaves et al. 2021]	X				
[Jalota and Agrawal 2021]	X	X			X

and fast processing. In [Gopalakrishnan et al. 2018], an objective was to identify the main factors impacting university student dropout. The work proposes tests with: *(i)* CS-filtering algorithms; *(ii)* information gain; *(iii)* correlation; *(iv)* relief; *(v)* maximum relevance and minimum redundancy (mRMR); and *(vi)* Kruskal Wallise test. Each algorithm produced a different attribute ranking result. The most important attributes evaluated included: the education level of the mother and the initial mathematical ability of a student. In [Febro 2019] the author employed some of the same filtering methods to show that family income and university entrance exam grade were the most significant factors for university student dropout in the Philippines.

In [Ramaswami and Bhaskaran 2009] the authors examined the how filtered FS techniques influence the predictive accuracy of student performance in India using demographical, socioeconomic, and school performance data. In [Dimic et al. 2019] the data from a grade management platform in an engineering faculty in Serbia was used to predict student performance using CS-filtering methods alongside relief feature scoring, and information gain.

Some works have also examined combinations of FS filters and classifiers to understand which one exhibited the best performance [Rachburee and Punlumjeak 2015; Abid et al. 2018; Enaro and Chakraborty 2020; Das et al. 2020; Muchuchuti et al. 2020]. In [Sokkhey and Okazaki 2020] the authors developed the CHIMI FS method, a combination of the CS and Mutual Information (MI) ranking algorithms. This approach was then used to identify the most relevant factors affecting student school performance. The work described in [Chaudhury and Tripathy 2020] proposes a two-staged FS technique. Initially, filter-based methods were used to analyze the student attributes that affect their academic performance. This was followed by the application of the Radial Bias Function Network (RBFN), a novel differential evolution algorithm that used to optimize the classification algorithm. Lastly, segregated instances of each class were used to identify the dominant features of the class.

Filtering methods were also combined with clustering-based approaches to identify students with

similar learning styles. This was performed in order to improve the prediction accuracy of student performance [Sasi Regha and Uma Rani 2015; Huang et al. 2019; Chaves et al. 2021]. In [Sasi Regha and Uma Rani 2015] the Non-negative Matrix Factorization Clustering based Feature Selection (NMFCFS) technique was introduced. NMFCFS uses Symmetric Uncertainty (SU) estimation for: *(i)* removing irrelevant features; and *(ii)* finding redundant features present in the relevant ones. The results showed that NMFCFS is able to achieve high accuracy performance when attempting to predict student failure and dropout. An entropy-based algorithm for finding important features for online learners clustering and analysis was proposed in [Huang et al. 2019]. The clustering algorithm k-means combined with the Correlation-based FS method was applied in [Chaves et al. 2021] to data originating from a physics course at a university in Spain. The results show that the method is able to clearly distinguish between students that pass and fail.

Wrapper approach uses a subset selection algorithm (*e.g.*, GA) as a wrapper around a classifier. This method was used in [Farissi et al. 2020] with a GA being used to construct the attribute sets to be tested. The main objective of the work was to classify student performance through demographic, behavioral, and academic formation attributes. The results obtained suggest an improvement in classifier performance by employing the features selected by the GA. In [Santos et al. 2020] the wrapper approach was also used with data from a grade management system from a Brazilian university. The dataset consisted of personal information combined with student academic performance. The authors proposed a Decision Tree (DT) classifier optimized with the help of a GA to predict student evasion risk.

An approach combining a GA alongside the K-Nearest Neighbors (KNN) classifier was utilized to predict school performance in [Wafi et al. 2019]. The method was able to achieve significantly better results when compared against a model that merely applied KNN. However, this improvement came at the expense of significant increases in processing time. In [Almasri et al. 2020] the authors evaluated 21 sampling algorithms in the wrapper approach to predict student performance in a Jordanian university. The best results were achieved with Bat Search, Harmony Search, and Ant Search.

Embedded approaches carry out FS as part of the building process of the classification model. In [Hassan et al. 2019] the authors proposed a Random Forest (RF) approach to assess student performance in an e-learning graduation of a Malaysian public university. The main attribute employed was the number of course visualizations. The method described in [Teodoro and Kappel 2020] also performed FS using a RF trained using a public database containing higher education data from Brazil. The main objective of the work was to determine the main characteristics leading to student dropout. The results obtained showed that the most relevant attributes for evasion were age, extracurricular activities, and total course load.

In [Gitinabard et al. 2018; Niu et al. 2018] DT ranking methods were employed to determine the main factors that result in students concluding free online courses. In [Gitinabard et al. 2018] the authors identified that watching videos is an important attribute for students to be able to gain the conclusion certificates offered in the online platforms *Coursera* and *EdX*. In [Niu et al. 2018], the authors concluded that the main attribute for predicting dropout in five multidisciplinary courses was the difference in the `Icourse163`[1] platform access times.

Feature extraction is also frequently used as a way to compact data. It transforms potentially redundant data into a new reduced set of features. In [Poudyal et al. 2020] a study was performed to predict student performance in higher education degrees using the Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) algorithms for feature extraction. The results obtained showed that the choice of which method to employ is dependent on the classifier utilized. The LDA technique produced good results for the logistic regression (LR) model. The PCA method achieved the best results when paired with the KNN classifier. Both approaches achieved similar

[1] Available at: `https://www.icourse163.com/`

results when combined with DTs. In [Šarić Grgić et al. 2020] a preprocessing stage was applied to perform a clustering analysis through k-means. The method performed a linear transformation of the attributes by using PCA. The main focus of the work was to identify patterns in an online tutoring platform of a university course.

Comparisons studies between different combinations of FS methods and classifiers have been performed in order to assess their respective individual performance. Namely, in [Ahmed et al. 2020] an evaluation study of student academic performance was executed using information from the department of computer science of a university in Bangladesh. In this work, the wrapper approach was combined with filtering techniques and paired alongside different classifiers such as KNN, Naive Bayes (NB), *Bagging*, RF, and DT. The combination which exhibited the best results was the one consisting of a GA in conjunction with KNN.

The approaches and results described in [Zaffar et al. 2018; Ajibade et al. 2019; Jalota and Agrawal 2021] used the same dataset of student academic performance in order to perform comparative analyses of different filtering techniques, wrappers, and extraction methods. More specifically, in [Zaffar et al. 2018] the authors evaluated several filtering methods with the feature extraction being performed by the PCA algorithm. In [Ajibade et al. 2019] a methodology was proposed with differential evolutionary algorithms that made use of sequential selection to construct the attribute sets. In [Jalota and Agrawal 2021] the NB model exhibited the best results when combined in the wrapper approach. The same work presented results where the DT was the best-suited model when the correlation filter was employed for FS.

In [Ahmed et al. 2019] the filter-based techniques, Pearson correlation and information gain, were combined with a wrapper method that uses a neural network (NN) as a baseline for comparing the effects of FS. This mechanism was evaluated by two datasets of student questionnaires. In [Govindasamy and Velmurugan 2019] a diverse set of feature extraction and filtering methods were evaluated for removing irrelevant data and their impact on various clustering algorithms was also assessed. The main objective of this work consisted in analyzing student performance by collecting questioners from various colleges. The approach described in [Hashemi et al. 2018] also used filtering and wrapper methods for FS. The main focus of the work was to predict university student acceptance based on results from a national exam in Iran. In [Punlumjeak and Rachburee 2015] FS was analyzed to predict academic performance in an engineering university in Thailand. The dataset utilized consisted of departmental data such as course grades. The proposal used the filter, wrapper and approaches with different classifiers.

The set of related works discussed above illustrates the importance that FS can have on algorithmic performance of machine learning methods. However, the works referenced only explore classical FS techniques without proposing an approach specific to educational data.

4. PROPOSED METHODOLOGY

The methodology adopted is presented in Figure 2 and consists of three main stages, namely: 1) data preprocessing; 2) FS through a GA with a flexible fitness function for educational data alongside CS and Correlation; and 3) classification.

The dataset employed contains the records of 480 students described through 16 attributes from an e-learning school platform[2] [Zaffar et al. 2018; Farissi et al. 2020; Jalota and Agrawal 2021]. The attributes are split amongst three main groups: (1) Behavioral characteristics: hand raised in class, number of site visits, number of news visualizations, engagement in forums, parental satisfaction surveys, degree of satisfaction, and number of absences; (2) Demographic characteristics: gender, place of birth and nationality; and (3) Academic education characteristics: educational stage, grade level,

[2]Available in the Kaggle repository: `https://www.kaggle.com/aljarah/xAPI-Edu-Data`

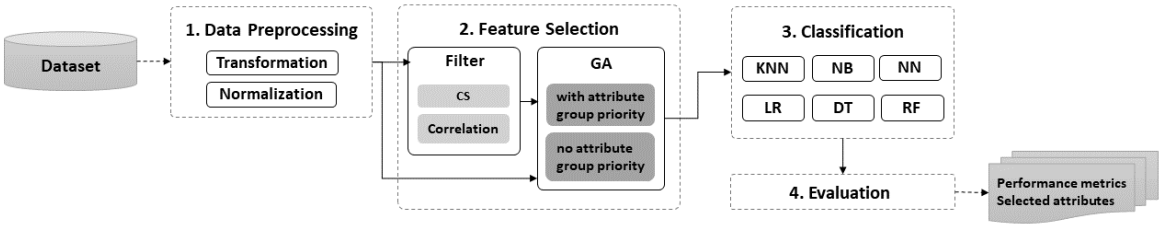

Fig. 2. Global perspective of the methodology adopted.

class, course, semester. Of the set of features available, the `Class` one represents student performance (low / average / high) and will also be the target attribute for classification. The percentages of low, average and high instance are, respectively, 29.37%, 26.45% and 43.90%.

The preprocessing stage includes the transformation of categorical data and the normalization of numerical data. All entries of the database detail the complete set of attributes, which is composed by 4 numerical features and 12 categorical ones. The categorical attributes are encoded into numerical ones. This procedure increases the number of features from 16 to 70. This stage was necessary to avoid considering categorical values as a numerical scale [Hancock and Khoshgoftaar 2020]. Data normalization was carried out through the MinMax technique as a way to avoid partial algorithms[3], with the values being altered to the range $[0, 1]$.

The next stage consists of an optimized selection of attributes for an educational context. The objective is to investigate if different subsets of attributes are more relevant than others when analyzing student performance. This information can then be used in a GA in order to prioritize the most relevant subsets. This requires that attributes need to be categorized into groups in accordance with their nature (behavioral, demographical, and academic).

The GA selects the attributes by creating subsets that will be evaluated by the OF proposed in Equation 1. Each subset is an individual represented by a binary vector b, where each element b_i indicates the presence (1) or the absence (0) of attribute i, and $|b| = N$ is the total number of attributes of the dataset. During each generation, the operations of crossover and mutation are applied. These are responsible for the evolutionary variation of each individual. The priorities of the attribute groups are set through weights in the OF to be maximized. The function is composed of three components: 1) classification accuracy with the selected attributes (acc); 2) a penalty referring to the number of attributes selected (num_select); and 3) the priority weights for each group of attributes, *i.e.*, the sum of the multiplication of the attribute weight i (w_i) by b_i.

$$OF = acc - \frac{num\_select}{N} + \frac{1}{\gamma} \times \sum_{i=1}^{N} w_i \times b_i \tag{1}$$

where $\gamma = \sum_{i=1}^{N} p_i$ is a normalization factor.

This approach provides education experts with a flexible tool that enables the attribution of priorities to the groups of features that are most relevant to a specific school environment. For instance, in face-to-face teaching, location attributes could be intuitively more relevant. As a result, a larger weight could be given to demographical attributes when compared to the one used to describe the importance of behavioral characteristics. The proposed GA was also incorporated into a hybrid approach that applies the classical filtering techniques (CS and Correlation) for attribute selection before

[3]`https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`

the GA. This was done as a way to maximize classification performance [Singh and Selvakumar 2015]. Furthermore, for comparison effects, a GA with an OF without attribute group weights was also examined.

In the final stage, the classifiers employed were, respectively: KNN, NB, NN, LR, DT, and RF. These classifiers were selected because they presented the most interesting results in the literature [Zaffar et al. 2018; Farissi et al. 2020]. KNN is a more traditional non-parametric technique for classification [Abu Amra and Maghari 2017]. It assumes that samples with the same class have similar characteristics. The distance between the new sampling and the others is calculated to determine its class. Although it is a simple technique with good results, it is a slow technique because it needs to explore all points in the training sample in the worst case [Abu Amra and Maghari 2017]. The NB algorithm is a probabilistic classifier based on Bayes' theorem. It assumes complete independence of the data features [de O. Santos et al. 2019], which enables a small amount of training data to be used in order to to estimate the parameters necessary for classification. A NN is a bio-inspired technique capable of modeling complicated non-linear relationships between input data and output of interest. This process is based on the learning process in human brains to determine classification rules. NN consists of many different layers of fundamental units (neurons) connected [Haykin 2004]. This work uses a Multi-Layer Perceptron (MLP) NN as described in [Haykin 2004]. LR is a supervised classification algorithm based on a linear model that maps the predicted values to probabilities using the Sigmoid function, which is a more complex cost function. It is a supervised classification algorithm [Hung et al. 2020]. DT is a supervised learning approach that predicts response values by learning decision rules from features [Murthy 1998; Sharma and Kumar 2016; Jalota and Agrawal 2019]. DT models are interpretable using "if-else" rules and treat categorical and continuous features in the same data set. The RF classification technique essentially builds a model consisting of a collection of DTs, each with its own unique set of attributes [Amrieh et al. 2016; Jalota and Agrawal 2019].

5. EXPERIMENTAL EVALUATION

The experiments were implemented in Python and executed in the Google Colab platform that was running on an Intel(R) Xeon(R) CPU @ 2.30GHz and had 13.3 GB of available RAM[4] using the libraries: *pandas*, *scikit-learn*, *prince* e *deap*[5]. The GA employed the following parameters: population = 30, generation = 30, *crossover* = 0.09 and mutation = 0.01. The KNN used euclidean distance with $k = 5$ and $p = 2$. The NN employed was the MLP with a maximum of 300 iterations. The NB utilized was gaussian with *smoothing* = 0.01. The RF used 15 estimators. The set of parameters employed was based on the choices of [Santos et al. 2020] and in parameter tests performed for that work. All the remaining GA and classifier parameters employed the default values defined in the libraries. The dataset was divided in the following manner: 70% for training and 30% for test. The f_1 metric was employed for evaluating model classification since it consists of an harmonic average of Precision (Equation 2) and Recall (Equation 3), as represented in Equation 4.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

where TP is true positive and FP is false positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

[4]Available in: https://github.com/SCICOM-CEFET-RJ/Educational-DataMining
[5]https://pandas.pydata.org, https://scikit-learn.org/stable, https://pypi.org/project/prince e https://deap.readthedocs.io/en/master

where FN represents the false negative values.

$$f1 = 2.0 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

In order to verify whether or not the differences of values obtained by the evaluated methods are statistically significant, a hypothesis test was executed. As it was not possible to prove that the data fit a normal distribution, the Willcoxon-Mann-Whitney nonparametric test was used [Fix and Jr 1955; Komatsu 2017]. For this test, the null hypothesis is that the distribution of the outcome is equal to the distribution of the same metric for another experiment. We used a 5% significance level, *i.e.*, the p-value found in the test should be at most 0.05 to reject the null hypothesis and conclude that two results are significantly different (*e.g.*, the result obtained without the use of FS and the one with the use of the method).

Initially, the GA was evaluated with a fitness function consisting of classifier accuracy and a penalty related to the number of attributes selected. The data showed that some attributes were more present regardless of the parameters and the classifiers utilized. As a result, we opted to evaluate attribute importance by ranking them through a filtering method[6]. The results showed that the attributes of the behavioral group are present in the first positions and that the ones referring to the academic characteristics are less prevalent. Thus, the attributes that reflect the attitudes of students inside the platform appear to be more relevant when defining their performance in a remote education scenario. Accordingly, prioritizing groups of attributes, more related to the context in which the students are a part, may help in the optimized selection of characteristics that contribute to the analysis of academic performance.

Based on this preliminary experiment, the proposed GA was evaluated by prioritizing the attributes of each group of the database (B: Behavioral, D: Demographical and A: Academic) and in the hybrid approach (consisting of a filtering method, CS and Correlation, and the GA). In an alternated manner, a group of attributes receives a weight of 0.7 whilst the remaining two are given the same priority, *i.e.* 0.15. Tables II and III present the results for the experiments: *(i)* without FS (baseline); *(ii)* GA with fitness function without group weights; *(iii)* GA with the fitness function proposed; and *(iv)* the hybrid approach (CS + GA; Correlation + GA) using the first 40 ranking attributes for both CS and Correlation filtering methods. For each classifier, the data presented consists of: *(i)* f_1 average and standard deviation for 10 executions and *(ii)* the average quantity of attributes selected. In order to support the analysis, we used the statistical test with a significance level of 5%. The underlined f_1 values are those obtained by the combination of methods that were statistically different and higher than those obtained by the baseline. Among these, the best for each combination are in bold (considering the significance level employed).

The KNN and NB classifiers presented performance gains when compared against the baseline that did not perform FS, as presented in Table II. In these cases, the GA approach consisting of priority groups produced the best average results when the largest weight was given to the behavioral attributes. In what concerns the hybrid approach with CS, the result with the behavioral group and the NN classifier was also more assertive on average than those obtained by the model that consider others priorities. In the context of the e-learning dataset employed, a flexible fitness function may be better suited to the different realities of the education scenario. Other classifiers, presented in Table III, showed only a statistically similar or lower f_1 score when compared against the baseline that did not perform FS.

Regarding the number of attributes selected: *(i)* the average value was smaller when the GA without priorities was tested; *(ii)* the number remained close to 35 for the GA proposed; and *(iii)* the value

[6]The first 40 attributes were selected with the largest CS ranking values. This choice was made since the set of 40 attributes presented larger f_1 values than the ones consisting of 10, 20 or 30 attributes.

Table II. Results for experiments using KNN, NB and NN classifiers

FS	KNN		NB		NN	
	f_1	# att.	f_1	# att.	f_1	# att.
without FS	67.1 ± 4.3	70	63.8 ± 4.6	70	72.6 ± 3.2	70
GA	70.8 ± 3.5	21	64.0 ± 4.3	22	72.4 ± 3.3	21
GA (B)	**77.2** ± 3.1	32	**69.0** ± 2.7	31	75.7 ± 2.4	32
GA (D)	73.2 ± 5.0	37	65.5 ± 3.0	34	74.0 ± 3.1	36
GA (A)	73.9 ± 3.5	35	66.0 ± 4.5	34	73.0 ± 4.2	38
CS + GA	70.9 ± 3.8	10	71.0 ± 4.0	9	71.1 ± 5.0	9
CS + GA (B)	76.2 ± 3.3	18	72.1 ± 2.4	18	**76.0** ± 1.8	17
CS + GA (D)	72.7 ± 4.0	21	67.0 ± 3.9	19	73.0 ± 3.8	22
CS + GA (A)	70.0 ± 6.8	19	70.8 ± 4.8	20	72.2 ± 3.7	20
Cor + GA	67.0 ± 5.8	10	70.7 ± 3.3	9	73.6 ± 3.0	9
Cor + GA (B)	75.5 ± 3.0	18	71.8 ± 2.4	17	75.4 ± 1.6	17
Cor + GA (D)	72.8 ± 3.7	22	66.5 ± 3.5	20	76.4 ± 3.2	21
Cor + GA (A)	72.3 ± 3.7	21	65.2 ± 5.0	21	73.3 ± 2.1	21

Table III. Results for experiments using LR, DT and RF classifiers

FS	LR		DT		RF	
	f_1	# att.	f_1	# att.	f_1	# att.
without FS	75.3 ± 2.7	70	72.6 ± 2.7	70	79.3 ± 2.9	70
GA	73.1 ± 2.8	21	68.7 ± 1.9	21	71.5 ± 3.6	21
GA (B)	75.3 ± 5.0	32	68.0 ± 2.8	33	76.4 ± 2.2	30
GA (D)	71.2 ± 2.9	36	68.9 ± 4.0	35	74.0 ± 3.1	34
GA (A)	74.6 ± 3.3	37	69.9 ± 4.2	37	71.2 ± 4.8	35
CS + GA	69.9 ± 3.2	9	67.6 ± 5.5	8	71.3 ± 4.0	9
CS + GA (B)	74.5 ± 3.5	18	71.6 ± 3.2	19	75.8 ± 2.3	18
CS + GA (D)	72.6 ± 2.5	22	68.1 ± 3.8	21	72.9 ± 4.0	22
CS + GA (A)	73.8 ± 3.3	20	65.0 ± 5.3	20	74.0 ± 2.8	21
Cor + GA	70.7 ± 5.2	9	60.7 ± 5.4	8	72.1 ± 3.5	10
Cor + GA (B)	76.1 ± 3.3	17	67.3 ± 3.3	18	75.7 ± 2.5	18
Cor + GA (D)	75.2 ± 2.6	20	69.5 ± 3.6	21	73.1 ± 3.2	22
Cor + GA (A)	73.3 ± 2.1	22	68.4 ± 2.3	21	71.9 ± 4.7	21

was 20 for the hybrid approach. The best f_1 score obtained with FS in the experiments was 77.2, and it was obtained in the GA (B) approach paired with KNN [Ahmed et al. 2020], which resulted in the selection of 32 attributes.

Our evaluation confirms that the proposed selection technique has an advantage in improving the classification of student performance with statistical significance for the KNN, NB and NN classifiers with greater weight for the behavioral attribute group. The result remains valid when both filter types are applied, which results in a reduction in the number of attributes.

6. FINAL REMARKS

The monitoring of students by educational professionals is an arduous task, involving numerous attributes. Feature selection is an important technique and has been increasingly used to identify the main factors that impact student performance in the context of EDM. This work presented a flexible objective function approach by extending the results of a previously published GA methodology. The main focus was on performing attribute selection in EDM in the context of e-learning. It is possible to state that the GA proposed increased, on average, the assertiveness of the FS procedure and, consequently, improved classification performance. In the e-learning education scenario, it was observed that the behavioral attributes are relevant to enhancing classification performance. However, in face-to-face education, other attributes, such as demographical ones, can be intuitively more pertinent. This dichotomy can be evaluated with the flexible approach presented. The approach

presented allows educational professionals to focus on the most important characteristics of each student. As future work, we intend to explore these techniques in other open educational databases in order to evaluate different educational environments, such as the one made available by the Brazilian government concerning higher education census.

Acknowledgements

This work was developed with the support of CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico (National Council for Scientific and Technological Development, in Brazil), CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Coordination for Enhancement of Higher Education Personnel, in Brazil), FAPERJ - Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (Rio de Janeiro Research Support Foundation, in Brazil), and CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (Federal Center for Technological Education of Rio de Janeiro).

REFERENCES

Abid, A., Kallel, I., Blanco, I., and Benayed, M. Selecting relevant educational attributes for predicting students' academic performance. *Advances in Intelligent Systems and Computing* vol. 736, pp. 650–660, 2018.

Abu Amra, I. A. and Maghari, A. Y. A. Students performance prediction using knn and naïve bayesian. In *2017 8th International Conference on Information Technology (ICIT)*. pp. 909–913, 2017.

Ahmed, M., Tahid, S., Mitu, N., Kundu, P., and Yeasmin, S. A comprehensive analysis on undergraduate student academic performance using feature selection techniques on classification algorithms. *11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*, 2020.

Ahmed, S., Al-Hamdani, R., and Croock, M. Edm preprocessing and hybrid feature selection for improving classification accuracy. *Journal of Theoretical and Applied Information Technology* 97 (1): 279–289, 2019.

Ajibade, S.-S., Ahmad, N., and Shamsuddin, S. An heuristic feature selection algorithm to evaluate academic performance of students. *ICSGRC 2019 - 2019 IEEE 10th Control and System Graduate Research Colloquium, Proceeding*, 2019.

Almasri, A., Alkhawaldeh, R., and Çelebi, E. Clustering-based emt model for predicting student performance. *Arabian Journal for Science and Engineering* 45 (12): 10067–10078, 2020.

Amrieh, E. A., Hamtini, T., and Aljarah, I. Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application* 9 (8): 119–136, 2016.

Chandrashekar, G. and Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering* 40 (1): 16–28, 2014. 40th-year commemorative issue.

Chaudhury, P. and Tripathy, H. A novel academic performance estimation model using two stage feature selection. *Indonesian Journal of Electrical Engineering and Computer Science* 19 (3): 1610–1619, 2020.

Chaves, V., Garcia Torres, M., Alonso, D., Gómez-Vela, F., Divina, F., and Vazquez Noguera, J. Analysis of student achievement scores via cluster analysis. *11th International Conference on EUropean Transnational Educational (ICEUTE 2020). Advances in Intelligent Systems and Computing* vol. 1266, pp. 399–408, 2021.

Chen, B., Hong, J., and Wang, Y. The minimum feature subset selection problem. *Journal of Computer Science and Technology* 12 (2): 145–153, 1997.

Chetana, V., Kolisetty, S. S., and Amogh, K. *A Short Survey of Dimensionality Reduction Techniques*. Recent Advances in Computer Based Systems, Processes and Applications, CRC Press, 2020.

Das, D., Shakir, A., Rabbani, M., Rahman, M., Shaharum, S., Khatun, S., Fadilah, N., Qaiduzzaman, K., Islam, M., and Arman, M. A comparative analysis of four classification algorithms for university students performance detection. *Lecture Notes in Electrical Engineering* vol. 632, pp. 415–424, 2020.

Dash, M. and Liu, H. Feature selection for classification. *Intelligent Data Analysis* 1 (1): 131 – 156, 1997.

Davies, S. and Russell, S. J. NP-completeness of searches for smallest possible feature sets. In *AAAI Symposium on Intelligent Relevance*. AAAI Press, pp. 37–39, 1994.

de Albuquerque, D., Brandão, D., and Coutinho, R. Um algoritmo genético com função de aptidão flexível para seleção de atributos em dados educacionais. In *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados*. SBC, Porto Alegre, RS, Brasil, pp. 355–360, 2021.

de O. Santos, K. J., Menezes, A. G., de Carvalho, A. B., and Montesco, C. A. E. 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. pp. 207–208, 2019.

Dimic, G., Rancic, D., Macek, N., Spalevic, P., and Drasute, V. Improving the prediction accuracy in blended learning environment using synthetic minority oversampling technique. *Information Discovery and Delivery* 47 (2): 76–83, 2019.

Enaro, A. and Chakraborty, S. Feature selection algorithms for predicting students academic performance using data mining techniques. *International Journal of Scientific and Technology Research* 9 (4): 3622–3626, 2020.

Farissi, A., Dahlan, H. M., and Samsuryadi. Genetic Algorithm Based Feature Selection for Predicting Student's Academic Performance. *Emerging Trends in Intelligent Computing and Informatics*, 2020.

Febro, J. Utilizing feature selection in identifying predicting factors of student retention. *International Journal of Advanced Computer Science and Applications* vol. 10, 01, 2019.

Fix, E. and Jr, J. L. H. Significance Probabilities of the Wilcoxon Test. *The Annals of Mathematical Statistics* 26 (2): 301 – 312, 1955.

Gitinabard, N., Khoshnevisan, F., Lynch, C., and Wang, E. Your actions or your associates? predicting certification and dropout in moocs with behavioral and social features. *Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018*, 2018.

Gopalakrishnan, A., Kased, R., Yang, H., Love, M., Graterol, C., and Shada, A. A multifaceted data mining approach to understanding what factors lead college students to persist and graduate. *Proceedings of Computing Conference 2017* vol. 2018-January, pp. 372–381, 2018.

Govindasamy, K. and Velmurugan, T. Preprocessing and feature extraction process in predicting students performance using clustering technique. *International Journal of Recent Technology and Engineering* 8 (1): 2407–2413, 2019.

Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. *Journal of Machine Learning Research 3 1157-1182*, 2003.

Han, J., Kamber, M., and Pei, J. *Data mining concepts and techniques, third edition*. Morgan Kaufmann Publishers, Waltham, Mass., 2012.

Hancock, J. and Khoshgoftaar, T. Survey on categorical data for neural networks. *Journal of Big Data* 28 (7), 2020.

Hashemi, H. Z., Parvasideh, P., Larijani, Z. H., and Morad, F. Analyze students performance of a national exam using feature selection methods. In *2018 8th International Conference on Computer and Knowledge Engineering (ICCKE)*. pp. 7–11, 2018.

Hassan, H., Anuar, S., and Ahmad, N. Students' performance prediction model using meta-classifier approach. *Communications in Computer and Information Science* vol. 1000, pp. 221–231, 2019.

Haykin, S. *Kalman filtering and neural networks*. Vol. 47. John Wiley & Sons, 2004.

Huang, L., Wang, X., Wu, Z., and Wang, F. Feature selection for clustering online learners. In *2019 Eighth International Conference on Educational Innovation through Technology (EITT)*. pp. 1–6, 2019.

Hung, H.-C., Liu, I.-F., Liang, C.-T., and Su, Y.-S. Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education. *Symmetry* 12 (2), 2020.

Jalota, C. and Agrawal, R. Analysis of educational data mining using classification. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. pp. 243–247, 2019.

Jalota, C. and Agrawal, R. Feature selection algorithms and student academic performance: A study. *Advances in Intelligent Systems and Computing* vol. 1165, pp. 317–328, 2021.

Kohavi, R. and John, G. H. Wrappers for feature subset selection. *Artificial Intelligence* 97 (1-2): 273–324, 1997.

Komatsu, A. Comparação dos poderes dos teste t de student e mann-whitney wilcoxon pelo método de monte carlo. vol. VI, pp. 121–127, 12, 2017.

Muchuchuti, S., Narasimhan, L., and Sidume, F. Classification model for student performance amelioration. *Lecture Notes in Networks and Systems* vol. 69, pp. 742–755, 2020.

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

Niu, Z., Li, W., Yan, X., and Wu, N. Exploring causes for the dropout on massive open online courses. In *Proceedings of ACM Turing Celebration Conference - China*. TURC '18. Association for Computing Machinery, New York, NY, USA, pp. 47–52, 2018.

Poudyal, S., Nagahi, M., Nagahisarchoghaei, M., and Ghanbari, G. Machine learning techniques for determining students' academic performance: A sustainable development case for engineering education. In *2020 International Conference on Decision Aid Sciences and Application, DASA 2020*. pp. 920–924, 2020.

Prabha, D., Siva Subramanian, R., Balakrishnan, S., and Karpagam, M. Performance evaluation of naive bayes classifier with and without filter based feature selection. *International Journal of Innovative Technology and Exploring Engineering* 8 (10): 2154–2158, 2019.

Punlumjeak, W. and Rachburee, N. A comparative study of feature selection techniques for classify student performance. *Proceedings - 2015 7th International Conference on Information Technology and Electrical Engineering: Envisioning the Trend of Computer, Information and Engineering, ICITEE 2015*, 2015.

Rachburee, N. and Punlumjeak, W. A comparison of feature selection approach between greedy, ig-ratio, chi-square, and mrmr in educational mining. *Proceedings - 2015 7th International Conference on Information Technology and Electrical Engineering: Envisioning the Trend of Computer, Information and Engineering, ICITEE 2015*, 2015.

Ramaswami, M. and Bhaskaran, R. A Study on Feature Selection Techniques in Educational Data Mining. *Journal of computing* 1 (1): 7–11, 2009.

Romero, C. and Ventura, S. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2010.

Santos, G. A., Belloze, K. T., Tarrataca, L., Haddad, D. B., Bordignon, A. L., and Brandao, D. N. EvolveDTree: Analyzing Student Dropout in Universities. *International Conference on Systems, Signals, and Image Processing* vol. 2020-July, pp. 173–178, 2020.

Sasi Regha, R. and Uma Rani, R. A novel clustering based feature selection for classifying student performance. *Indian Journal of Science and Technology* vol. 8, pp. 135–140, 2015.

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

Singh, S. and Selvakumar, S. A hybrid feature subset selection by combining filters and genetic algorithm. In *ICCCA*. pp. 283–289, 2015.

Sokkhey, P. and Okazaki, T. Study on dominant factor for academic performance prediction using feature selection methods. *International Journal of Advanced Computer Science and Applications* 11 (8): 492–502, 2020.

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

Teodoro, L. d. A. and Kappel, M. A. Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no brasil. *Revista Brasileira de Informática na Educação* 28 (0): 838–863, 2020.

Ullah, A., Khan, F. H., Qamar, U., and Bashir, S. dimensionality reduction approaches and evolving challenges in high dimensional data. *ACM International Conference Proceeding Series*, 2017.

Velliangiri, S., Alagumuthukrishnan, S., and Thankumar joseph, S. I. A review of dimensionality reduction techniques for efficient computation. *Procedia Computer Science* vol. 165, pp. 104–111, 2019.

Venkatesh, B. and Anuradha, J. A review of feature selection and its methods. *Cybernetics and Information Technologies* 19 (1): 3–26, 2019.

Wafi, M., Faruq, U., and Supianto, A. Automatic feature selection for modified k-nearest neighbor to predict student's academic performance. *Proceedings of 2019 4th International Conference on Sustainable Information Engineering and Technology, SIET 2019*, 2019.

Wang, S., Tang, J., and Liu, H. Embedded unsupervised feature selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29, 2015.

Zaffar, M., Hashmani, M. A., and Savita, K. S. Performance analysis of feature selection algorithm for educational data mining. *2017 IEEE Conference on Big Data and Analytics, ICBDA 2017* vol. 2018-January, pp. 7–12, 2018.

Šarić Grgić, I., Grubišić, A., Šerić, L., and Robinson, T. Student clustering based on learning behavior data in the intelligent tutoring system. *International Journal of Distance Education Technologies* 18 (2): 73–89, 2020.