Educational Data Mining: Discovery Standards of Academic Performance by Students in Public High Schools in the Federal District of Brazil

Eduardo Fernandes[1,2(✉)], Rommel Carvalho[1,3], Maristela Holanda[1], and Gustavo Van Erven[1,3]

[1] Department of Computer Science (CIC), University of Brasilia (UnB), Brasilia, DF, Brazil
eduardo.fernandes@se.df.gov.br, mholanda@cic.unb.br
[2] Subsecretariat for Modernization and Technology (SUMTEC), Education Secretary of State of the Federal District (SEDF), Brasilia, DF, Brazil
[3] Department of Research and Strategic Information (DIE), Ministry of Transparency, Monitoring and Control (MTFC), Brasilia, DF, Brazil
{rommel.carvalho,gustavo.erven}@cgu.gov.br

Abstract. This article presents results obtained in research regarding the academic performance of high school students at public schools in the Federal District of Brazil in 2015. Using CRISP-DM data mining methodology, we were able to achieve greater knowledge discovery than studies using traditional descriptive statistical analysis. Subsequently, our data shows that the variables, 'grades' and 'absences', are not the only attributes relevant to whether a student will fail at the end of the school year. Thus, this study presents data indicating other frequently reported attributes relevant to potential academic failure in this context, as well as a detailed explanation of the methodology, and the steps taken to obtain this data.

Keywords: Educational data mining · Academic performance · Data science · H2O · CRISP-DM · GBM · Decision Tree

1 Introduction

Descriptive statistic is the selection, analysis, and interpretation of numerical data through the elaboration of adequate: charts, graphs, and numerical indicators [1]. In other words, descriptive statistics can be considered as a set of analytical techniques used to summarize data collected in a given investigation, which, in the case of this paper, refers to data on selected variables regarding high school students from public schools in the Federal District of Brazil in 2015. These are usually organized as figures, tables, and graphs, with the intention of providing reports to submit information on the central tendency and dispersion of data. Therefore, the usual parameters described are: minimum value, maximum value, sum of the values, scores, mean, mode, median, variance, and standard deviation.

Á. Rocha et al. (eds.), *Recent Advances in Information Systems and Technologies*, Advances in Intelligent Systems and Computing 569, DOI 10.1007/978-3-319-56535-4_29

Descriptive statistical analysis is effective in providing basic descriptive information of a specific set of data. However, it does not facilitate the discovery of new patterns of knowledge related to the set of data. Therefore, this paper aims to answer the following research question: how can new, implicit data discovery standards be obtained, in addition to the information already obtained by descriptive statistical analysis, from the grades of high school students in the public schools of the Federal District of Brazil?

This paper aims to help students in the public schools of the Federal District of Brazil, who are in their third year of high school, improve their grades and avoid failing. The data mining classification method, and the algorithm Gradient Boosting Machine (GBM) [18], were used at the beginning of the year to map the most relevant variables in indicating low achievement and failure, even before having any of the students' grades. Subsequently, over time, we incorporated scores as they became available - their bimonthly grades - to verify how they impact the precision of the model, and help identify students who are prone to failure. Subsequently, the pedagogical support given to these students can be made more efficient, resulting in a lower failure rate. As a specific focus, we use the data of the students in their third year of high school to find out who needs educational support during the year, thereby minimizing failure rates at the end of the year.

Thus, this paper is organized as follows. Section 2 describes a literature review of Educational Data Mining. Section 3 describes the related works. Section 4 describes the metodology utilized. Section 5 presents the results. Finally, Sect. 6 outlines the conclusions.

2 Educational Data Mining

The mining of educational data can be defined as the application of techniques of traditional data mining on educational data analysis aimed at solving problems in the educational context [3]. To understand more about students, their academic performance, learning styles, and various issues directly linked to these, researchers are developing data mining methods to explore this context [4].

Currently, many states already have a large amount of data related to the academic progress of students in various types of schools. The collection of this data is the result of the modernization of data collecting instruments in the area of education, with various educational softwares and school management instruments now available [5]. For example, the State Secretary of Education of the Federal District has been using the free software iEducar[1] managing of information on students in the school.

[1] According to the Brazilian public software portal [6]. The iEducar software aims to centralize the information of a school system, which may be local, state, or even federal, depending on customizations that are possible within the system to suit each of their specific needs. Besides this main purpose, iEducar was designed to use less paper, eliminating the need to duplicate documents, and reducing the time needed to respond to requests, thus streamlining the work done by public workers.

Educational data mining facilitates, for example, the discovery of new patterns and knowledge about the process of student learning. Using this model, we can validate and evaluate some aspects of the educational system with the goal of improving the quality of education [7]. Some of these ideas sprang from the application of data mining in e-commerce systems that aim to identify consumer interests with the purpose of improving sales [8]. However, there are some points that differentiate educational data mining from traditional data mining, such as: objectives, the dataset, and techniques [8].

Meanwhile, although most traditional data mining techniques can be applied directly to educational data, some need adjustments to achieve their purpose with educational data. The educational environment has several natural groups, such as students, teachers, coordinators, and directors. Thus, the educational information may be analyzed from different angles, since each group has its own mission and goals [9]. As a practical example, the discovery of new patterns in student learning can be used by teachers to prepare their lessons. Students also benefit by getting feedback about their own learning process [10].

Notably, most studies in the 1990s also aimed at predicting student performance, even with a much smaller amount of data than available today. The current work of Cristobal et al. [2], published in magazines and shared at educational conferences, with the new data generated by information systems in educational environments is even more relevant. In addition to having a higher degree of confidence with the new techniques, there has never before been so much educational data as has been obtained in recent years.

3 Related Work

In the educational environment, various tasks and applications have been achieved using data mining. For example, Baker et al. [4,11] suggested four major areas of application for educational data mining: improving student models; improving the domain model; studying pedagogical support using learning software; and scientific research on student learning. In addition, they have also suggested five methods: prediction; clustering; relationship mining; data distillation for human judgment; and discovery with models.

Castro et al. [12] suggests the following tasks for educational data mining: applications that deal with the evaluation of students' academic performance; applications that provide adaptive courses according to the students' learning behavior; applications that evaluate educational resources available on Web courses; applications involving feedback to teachers and students in distance education courses; and applications that address atypical behaviors of student learning.

Romero et al. [2] established their own categories for the main tasks that make use of techniques for educational data mining: Analysis & Visualization, Providing Feedback, Recommendation, Predicting Performance, Student Modeling, Detecting Behavior, Grouping Students, Social Network Analysis, Developing Concept Map, Planning & Scheduling and Constructing Courseware. The most commonly

used tasks are regression, classification, clustering, and association rules. The most used algorithms are the decision trees, neural networks and Bayesian networks.

In this paper we present a model to evaluate the performance of third year high school students, as early in the year as possible, with the view to map out policies and carry out educational interventions in due time, in order to improve the students' performance at the end of each school year, minimizing the number of failures.

4 Methodology

Given that this is a data mining project, the most suitable method for achieving the expected results is the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology [15], which is the most common methodology for data mining, with a preference rate among professionals of 42% [13]. The selection of this methodology in lieu of others, such as SEMMA (Sample, Explore, Modify, Model, Assess) [14] or KDD (Knowledge Discovery in Databases) [14], is based on the fact that CRISP-DM is the most complete, and that it starts with the important phase of business awareness. The CRISP-DM is a methodology that focuses on the needs of managers to solve their management problems. It is the methodology which consists of a cycle that features six steps, as shown in Fig. 1 [15]:

1. *Business Understanding*. This initial phase focuses on understanding the project objectives and requirements from a business perspective.
2. *Data Understanding*. The data understanding phase starts with an initial data collection and proceeds with activities designed to familiarize users with them;

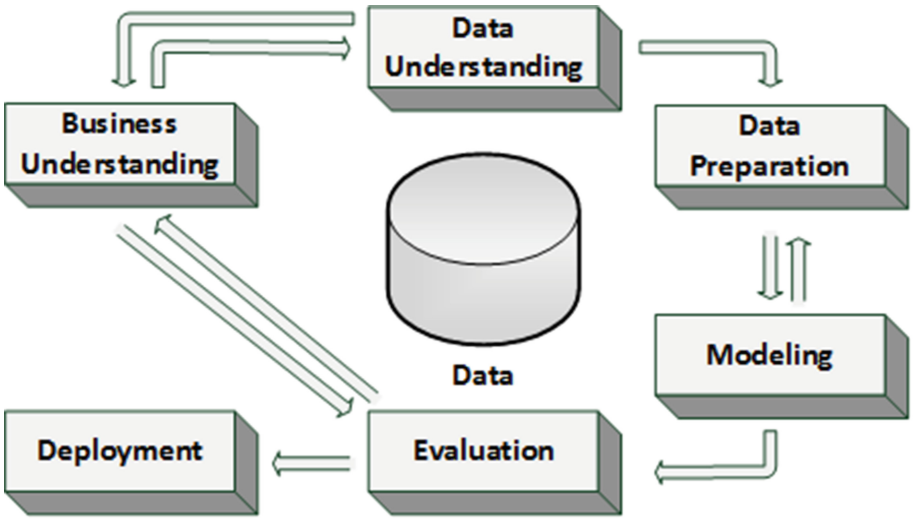

Fig. 1. The CRISP-DM methodology.

3. *Data Preparation.* The data preparation phase covers all activities to build the final set of data (data that will be fed into the modeling tool) from the initial raw data.
4. *Modeling.* In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values.
5. *Evaluation.* This phase of the project is to develop high quality models from the perspective of data analysis, and to evaluate if the generated model solves the problems raised in the Business Understanding phase.
6. *Deployment.* Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing the data mining process throughout the enterprise.

Since it is a reference model, the CRISP-DM is thorough and documented. All phases are organized, structured, and defined, so that the project is easily understood and revised [16].

In this paper, was used all the CRISP-DM phases except the deployment phase, which will be defined in future works.

5 Results

In this section, we will discuss the results of this paper according to each phase of the CRISP-DM.

In the **Business Understanding** phase, the data for doing this work were taken from the iEducar database, which is used in the State Department of Education of the Federal District in Brazil. The data obtained have several attributes related to each student, such as bimonthly grades, courses taken, and absences, among others. After retrieving this dataset, the research objective was defined - to decrease the annual retention rate of students in their third year of high school. This is achieved by performing the classification task at the beginning of the year, even before the first grades are calculated, so that students who have the highest probability of failing at the end of the year are identified in order to concentrate mentoring efforts on these students.

To carry out the **Data Understanding** phase, the database was imported on H2O Flow [17], which is described as a notebook-style open-source user interface for H2O. Using this interface we can import files, build models, and iteratively improve them, as well as work with a large amount of data in parallel, to generate better results in less time. Based on our models, we can make predictions and add rich text to create vignettes of our work – all within the Flow's browser-based environment.

After this importation, we performed a simple descriptive analysis on 17 variables: the regional name of education, the name of the city where the school is based, the school name, the time the class is given, whether the class has a student with a disability, the type of classroom, the student's name, sex, age, government grant status, student's place of residence - city, neighborhood, indication of special needs if any, courses taken, grades for the first two months,

absences, and whether the student has been approved or not at the end of the school year. This descriptive analysis has generated the domain of each variable (how many records and which), and the means, variances, mode, and standard deviation.

At this stage, the major findings of the 238,575 records in the database indicate that among the 18,908 students in their third year of high school, in 86 schools, 14 in regional education, participants in this research failed at a rate of 12,41% in the year 2015. As the purpose of this paper is to present findings that may help to reduce this failure rate, these data were essential in setting goals for the coming year. Moreover, it was found that these students reside in 46 different cities, both within and outside the Federal District and are on the average 16.89 years old.

At the **Data Preparation** stage, first we selected which variables of the set would be used to generate the first model, i.e., for identifying the probability of failure at the end of the year based on information only available at the beginning of the year. Therefore, these variables could not contain grades, subjects and absences from the first two months. Moreover, the students' names were not included, since they function only as an identification of that student, and it does not facilitate identifying patterns. These variables were used for creating the first model. For the second model, which aims to identify the probability of failure at the end of the year, only after the first two months, the variables previously ignored were then included (except the students' names).

In the **Modeling** phase we chose the Gradient Boosting Machine (GBM) [18] Algorithm, because it is a classification algorithm that produces a predictive model in the form of a set of weak prediction models, known as decision trees, which provide a rate for the importance of each attribute of our database to determine whether that student will pass at the end of the school year. The implementation of the algorithm in H2O was chosen because it is able to parallelize the task classification according to the machine's number of scores, facilitating the processing of a greater amount of data and providing much faster results. The Boosting is a flexible nonlinear regression procedure that helps improve the accuracy of trees [18]. By sequentially applying weak classification algorithms to incrementally changing data, a series of decision trees are created that produce an ensemble of weak prediction models.

With the initial base, in 2015, there was a failure rate of 12.41%, as seen in the data understanding phase. After the generation of the first model, at the **Evaluation** stage, we found that the training data - ROC curve - was 0.919, as can be seen in Fig. 2(a). With regard to the validation data, our ROC curve was 0.908, as can be seen in Fig. 2(b). These measurements show that there was a good rate in the sensitivity of the generated model successes, shown in the matrix of confusion according to Table 1, which shows the amount of records in agreement with regard to hits per class. We can also check the degree of importance of each variable to generate the data in Fig. 3.

After generating the second model, which adds information from the end of the first academic quarter, we found that the training data ROC curve was 0.950,

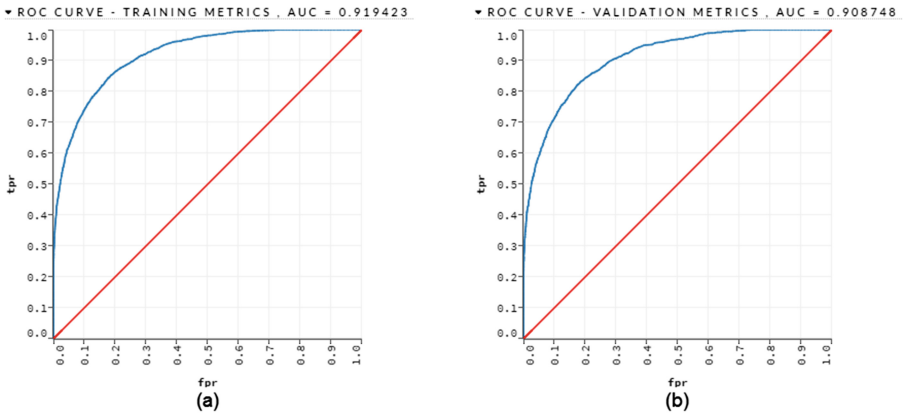

Fig. 2. ROC curve - beginning of the year - (a) Training metrics and (b) Validation metrics.

Table 1. Confusion matrix - beginning of the year.

	Approved	Disapproved	Error rate
Approved	199,241	9,472	0.045383 = 9472/208713
Disapproved	11,948	17,914	0.400107 = 11948/29862
Totals	211,189	27,386	0.089783 = 21420/238575

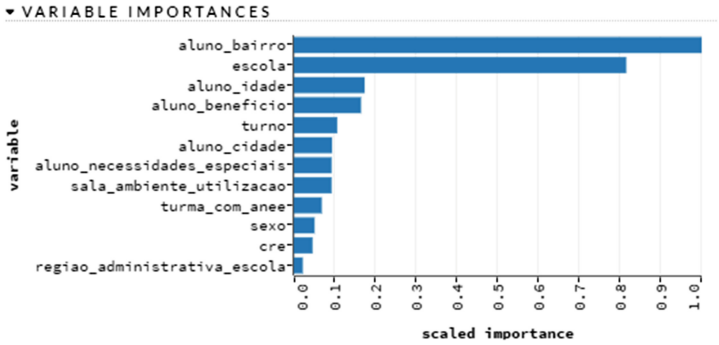

Fig. 3. Variable importance - beginning of the year.

shown in Fig. 4(a) and the validation data ROC curve was 0.913, shown in Fig. 4(b). These measurements establish that there was a good rate in the sensitivity of the hits generated model, as we see in the confusion matrix presented in Table 2, which displays the amount of hits per class. We can also identify the degree of importance of each variable to generate the data, shown in Fig. 5.

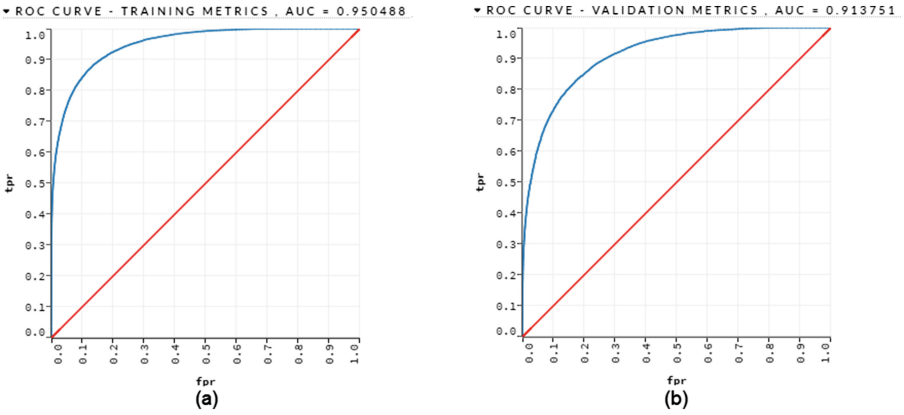

Fig. 4. ROC curve - after 1 bimester - (a) Training metrics and (b) Validation metrics.

Table 2. Confusion matrix - after 1 bimester.

	Approved	Disapproved	Error rate
Approved	200,014	8,699	0.041679 = 8699/208713
Disapproved	9,969	19,893	0.333836 = 9969/29862
Totals	209,983	28,592	0.078248 = 18668/238575

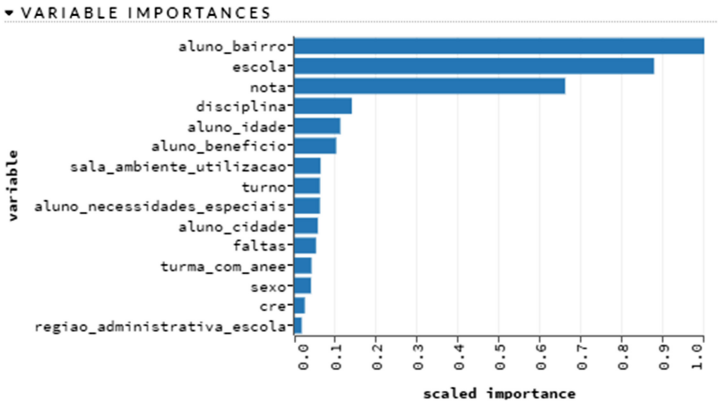

Fig. 5. Variable importance - after 1 bimester.

We can conclude that by identifying the students' subjects, grades and absences in the first two months, there is greater assertiveness in discovering which students may eventually fail. Identifying the degrees of importance of each variable in the model, and comparing these to the first, the variable, "grades", was shown to rank in third place, preceded only by the students' neighborhood, and his or her school.

6 Conclusion

Aiming to aid students in their third year of high school in the public schools of the Federal District of Brazil, this research has identified a series of variables that may predict the probability of success or failure of these students, early in the school year, so that educators may design interventions that will assure their success.

We used the method of data mining classification, with GBM algorithm, which indicated the attributes identified at the beginning of the school year that most contributed to failure rates, without necessarily taking into consideration the students' grades and the students' number of absences, which are not available at the beginning of the year.

The standard found in this paper was that the variables neighborhood and school where the students study are the main factors that influence students' failure. Accuracy in determining the failure rate increased over time, with the availability of bimonthly grades, and the number of absences to add to the model. With the research results obtained and presented in this paper, pedagogical support for students can be more efficient, and help students who have the greatest need, so that they may pass at the end of the school year, thereby reducing the failure rate.

The next steps proposed for this study are the deployment of the model generated into the State Department of Education of the Federal District. In addition, we aim to create models that include grades and the number of absences from other semesters to demonstrate that with the insertion of these data, the model tends to be more precise, thus, academic advising can be considerably more effective throughout the year.

References

1. Reis, E.: Estatstica Descritiva, 245 p. Edies Slabo Lda, Lisboa (1991). ISBN 972-618-060-0
2. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. **40**(6), 601–618 (2010)
3. Barnes, T., Desmarais, M., Romero, C., Ventura, S.: Presented at the 2nd International Conference on Educational Data Mining, Cordoba, Spain (2009)
4. Baker, R.: Data mining for education. In: McGaw, B., Peterson, P., Baker, E. (eds.) International Encyclopedia of Education, 3rd edn. Elsevier, Oxford (2010)
5. Koedinger, K., Cunningham, K., Skogsholm, A., Leber, B.: An open repository and analysis tools for fine-grained, longitudinal learner data. In: Proceedings of the 1st International Conference on Educational Data Mining, Montreal, QC, Canada, pp. 157–166 (2008)
6. https://softwarepublico.gov.br/social/i-educar. Accessed July 2016
7. Romero, C., Ventura, S., De Bra, P.: Knowledge discovery with genetic programming for providing feedback to courseware author. User Model. User-Adap. Inter.: J. Personalization Res. **14**(5), 425–464 (2004)
8. Raghavan, S.N.R.: Data mining in e-commerce: a survey. Sadhana J. **30**(2/3), 275–289 (2005)

9. Hanna, M.: Data mining in the e-learning domain. Campus-Wide Inf. Syst. **21**(1), 29–34 (2004)
10. Merceron, A., Yacef, K.: Educational data mining: a case study. In: Proceedings of the International Conference on Artificial Intelligence in Education, Amsterdam, The Netherlands, pp. 1–8 (2005)
11. Baker, R., Yacef, K.: The state of educational data mining in 2009: a review and future visions. J. Educ. Data Mining **1**(1), 3–17 (2009)
12. Castro, F., Vellido, A., Nebot, A., Mugica, F.: Applying data mining techniques to e-learning problems. In: Jain, L.C., Tedman, R., Tedman, D. (eds.) Evolution of Teaching and Learning Paradigms in Intelligent Environment. SCI, vol. 62, pp. 183–221. Springer, Heidelberg (2007)
13. http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html. Accessed July 2016
14. Azevedo, A., Santos, F.: KDD, SEMMA and CRISP-DM: a parallel overview. In: IADIS European Conference Data Mining (2008)
15. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0 step-by-step data mining guide. SPSS (2000)
16. Santos, M., Azevedo, C.: Data Mining: Descoberta de Conhecimento em Bases de Dados. FCA ed., Lisboa (2005)
17. http://www.h2o.ai/product/flow/. Accessed July 2016
18. http://www.h2o.ai/verticals/algos/gbm/. Accessed July 2016