ARTICLE IN PRESS

Journal of Business Research xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

JOURNAL OF BUSINESS RESEARCH

Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil

Eduardo Fernandes[a,b,*], Maristela Holanda[a], Marcio Victorino[a], Vinicius Borges[a], Rommel Carvalho[a], Gustavo Van Erven[a,c]

[a] Department of Computer Science (CIC), University of Brasilia (UnB), Brasilia, DF, Brazil
[b] Subsecretariat for Modernization and Technology (SUMTEC), Education Secretary of State of the Federal District (SEDF), Brasilia, DF, Brazil
[c] Department of Research and Strategic Information (DIE), Ministry of Transparency, Monitoring and Control (MTFC), Brasilia, DF, Brazil

ARTICLE INFO

Keywords:
Educational data mining
Academic performance
Predictive analysis
Decision tree
Gradient boosting machine
H2O

ABSTRACT

In this article, we present a predictive analysis of the academic performance of students in public schools of the Federal District of Brazil during the school terms of 2015 and 2016. Initially, we performed a descriptive statistical analysis to gain insight from data. Subsequently, two datasets were obtained. The first dataset contains variables obtained prior to the start of the school year, and the second included academic variables collected two months after the semester began. Classification models based on the Gradient Boosting Machine (GBM) were created to predict academic outcomes of student performance at the end of the school year for each dataset. Results showed that, though the attributes 'grades' and 'absences' were the most relevant for predicting the end of the year academic outcomes of student performance, the analysis of demographic attributes reveals that 'neighborhood', 'school' and 'age' are also potential indicators of a student's academic success or failure.

1. Introduction

Descriptive statistics is the selection, analysis, and interpretation of numerical data through the elaboration of adequate: charts, graphs, and numeric indicators (Reis, Melo, Andrade, & Calapez, 1999). 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 with respect to high school students in public schools of the Federal District of Brazil, during 2015 and 2016. These are typically organized as figures, tables, and graphs, to provide reports to submit information on the central tendency and dispersion of data. Thus, the parameters described here are: minimum value, maximum value, sum of the values, scores, mean, mode, median, variance, and standard deviation.

Descriptive statistical analysis is effective in providing basic descriptive information of a specific dataset. However, the discovery of new patterns of knowledge related to the mass of data is costly. Therefore, in an effort to deal with this mass of data and discover new patterns more efficiently, this paper aims to answer the following research question: how can we obtain implicit patterns of discovery data for public high school students in the Federal District of Brazil that goes beyond observing the information already provided by descriptive statistical analysis comprised exclusively of their grades?

The answer to this question and the subsequent knowledge acquired supports the general goal of this study of providing information to city officials, teachers and guidance counselors, which will aide them in the development of public policies, didactic materials and social work initiatives, in order to support students in the public schools of the Federal District of Brazil. The target group are students in their third year of high school, and our specific goal is to help them improve their grades to at least a passing level so that their academic trajectory follows uninterrupted, guaranteeing their graduation, and a possibility of higher education.

To achieve these goals, the data mining classification method, and the algorithm Gradient Boosting Machine (GBM)[1], were used at the beginning of 2015 and 2016, 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 their achievement scores as they became available – their bimonthly grades – to verify how these grades may have an impact on the precision of the model, and to help identify students who were prone to failure. Ultimately, the pedagogical support given to these students becomes more

* Corresponding author at: Department of Computer Science (CIC), University of Brasilia (UnB), Brasilia, DF, Brazil.
E-mail addresses: eduardo.fernandes@se.df.gov.br (E. Fernandes), mholanda@cic.unb.br (M. Holanda), mcvictorino@unb.br (M. Victorino), rommel.carvalho@cgu.gov.br (R. Carvalho), gustavo.erven@cgu.gov.br (G.V. Erven).

[1] H2o gbm, https://www.h2o.ai/h2o/machine-learning/, accessed: 2017, July.

https://doi.org/10.1016/j.jbusres.2018.02.012
Received 15 August 2017; Received in revised form 6 February 2018; Accepted 7 February 2018

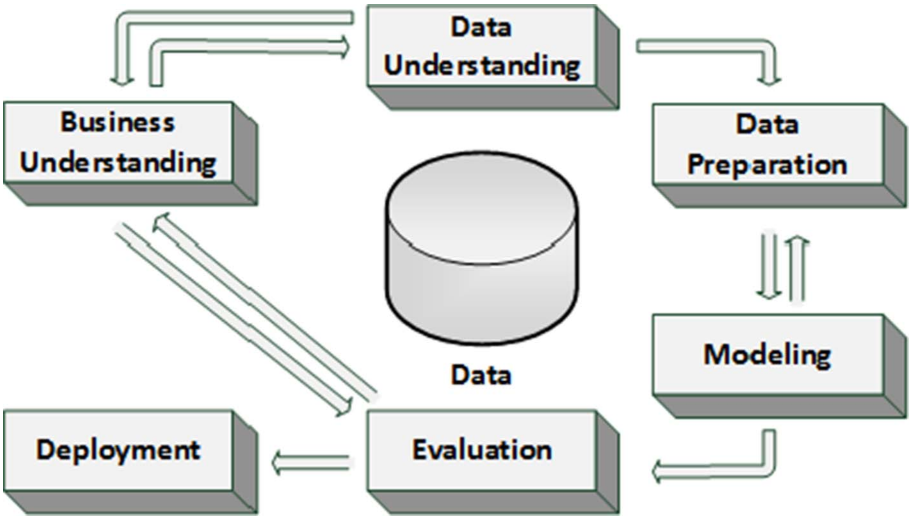

Fig. 1. The CRISP-DM methodology.

efficient with this information, specifically, data regarding third year high school students, indicating at the beginning of the year which of them may need educational support, thereby minimizing failure rates at the end of the year.

Thus, this paper is organized as follows. Section 2 presents a literature review of Educational Data Mining. Section 3 describes the related works. Section 4 describes the methodology utilized. Section 5 presents the results. Finally, Section 6 outlines the conclusions and indicates possibilities for future work.

2. Educational data mining

Educational Data Mining (EDM) can be defined as the application of techniques of traditional data mining to educational data analysis aimed at solving problems in the educational context (R. S.~Baker & Yacef, 2009). Some EDM applications comprise the development of e-learning systems (Lara, Lizcano, Martínez, Pazos, & Riera, 2014), pedagogical support (Hung & Crooks, 2009), clustering educational data (Chakraborty, Chakma, & Mukherjee, 2016), student performance predictions (Kabra & Bichkar, 2011). This research focuses on the student performance predictions since we are interested in understanding the main factors (social, personal and academic) that affect their performance at school.

To better understand students' academic performance and learning styles, as well as various issues directly linked to these, researchers have been developing data mining methods that explore this context (R. S. J. D.~Baker, 2010). Currently, many States in Brazil 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 field of education, with various educational softwares and school management instruments (Koedinger, Cunningham, Skogsholm, & Leber, 2008). For example, the State Secretary of Education of the Federal District has been using the free software iEducar[2] managing student information.

Table 1
Comparative analysis between 2015 and 2016.

	2015	2016	Type of variable
Regional coordination education	14	14	Nominal
School administrative region	24	24	Nominal
School	86	86	Nominal
Shift	4	4	Nominal
Class with persons with special needs	2	2	Nominal
Classroom usage environment	17	2	Nominal
Student code	19,000	19,834	Nominal
Gender	2	2	Nominal
Student age (mean)	16.8923	16.8666	Numeric
Student benefit	5	4	Nominal
Student city	44	46	Nominal
Student neighborhood	301	305	Nominal
Student with special needs	42	39	Nominal
School subjects	17	17	Nominal
Grade (mean)	5.8467	5.9247	Numeric
Absence (mean)	1.9559	2.0166	Numeric
Student end result	2	2	Nominal

The EDM uses the databases of these educational systems to understand the students and their learning styles more comprehensively in an effort to design educational policies that will improve their academic performance, and reduce failure rates at the end of each school year. The EDM facilitates, for example, the discovery of new patterns and knowledge about the process of student learning. Using this model, one can validate and evaluate some aspects of the educational system with the goal of improving the quality of education (Romero, Ventura, & De~Bra, 2004). Some of these ideas sprang from the application of data mining in e-commerce systems that aim to identify consumer interests, which would potentially help improve sales. However, there are some points that differentiate EDM from traditional data mining, such as: objectives, the dataset, and techniques (Raghavan, 2005). Meanwhile, although most traditional data mining techniques can be applied directly to educational data, some need adjustments to achieve their purpose, due to the specificities of the educational environment. This 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 (Hanna, 2004). Ultimately, this information can have practical use for these groups, for 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 (Merceron & Yacef, 2005). It is worthy to note that the majority of studies in the 1990s also aimed at predicting student performance, but with a much smaller amount of data than is available today. Subsequently, the more recent work of researchers, such as, Romero and

[2] According to the Brazilian public software portal (Portal do software público brasileiro – I-Educar, n.d.). The iEducar software aims to centralize the information of particular school systems, which may be local, state, or even federal, depending on customizations that are possible within the system that suit each of their specific needs. In addition to this central purpose, iEducar was designed to use less paper, eliminating the need for duplicating documents, and reducing the time needed to respond to requests, thus streamlining the work done by public workers.

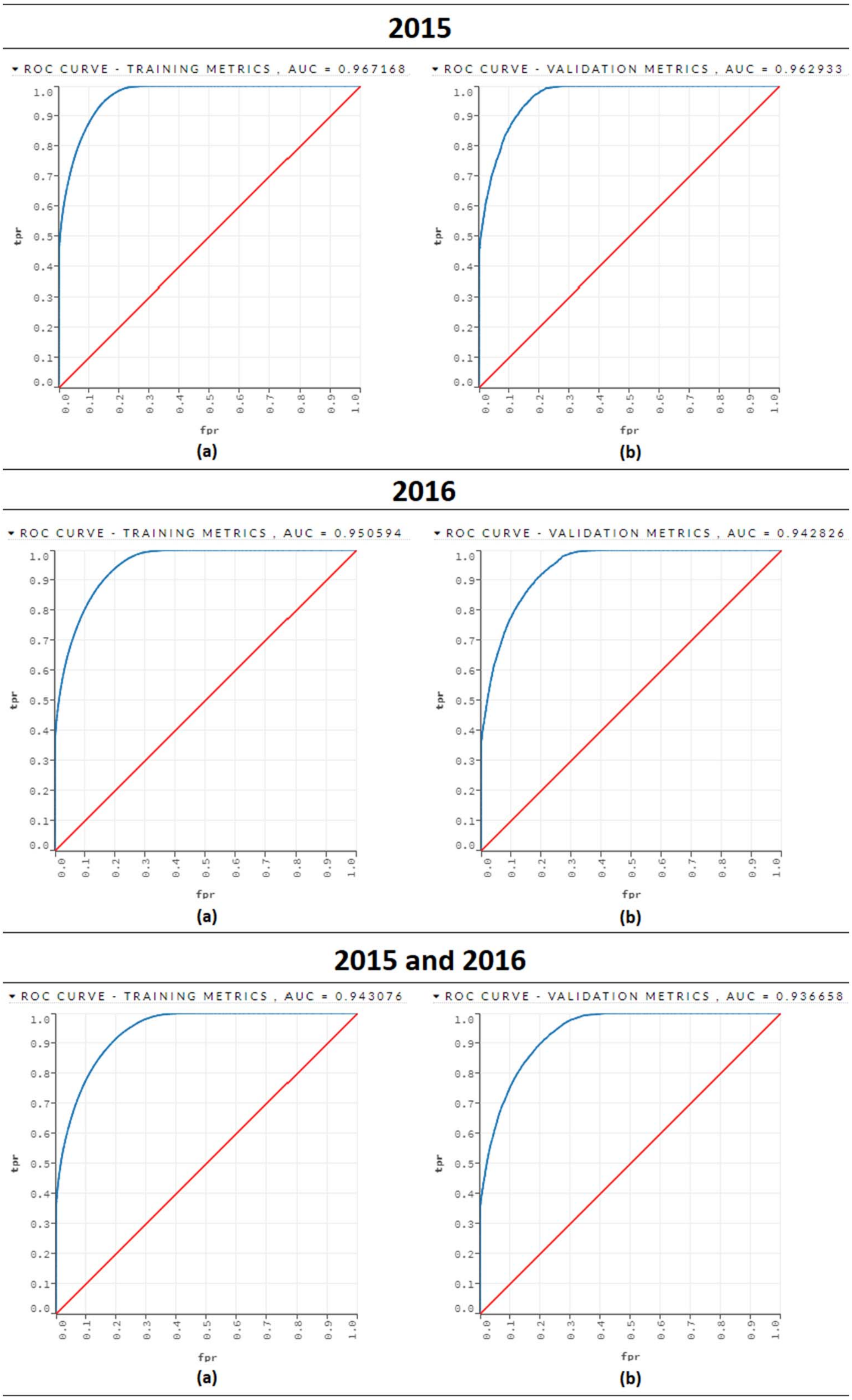

Fig. 2. ROC curve for CM-I - beginning of the year - (a) training metrics and (b) validation metrics.

Ventura (2010), are now able to provide more relevant and reliable information with the new data generated by information systems in educational environments. In addition to providing a higher degree of confidence, these new techniques are also able to manage the massive amounts of educational data obtained in recent years.

3. Related work

In the educational environment, data mining has been applied for various purposes and to accomplish various tasks. For example, R. S. J. D.~Baker (2010) and R. S.~Baker and Yacef (2009) suggested four

Table 2
Confusion matrix for CM-I - beginning of the year.

	Approved	Disapproved	Error	Rate
2015				
Approved	194,547	14,166	0.067873	=14,166/208,713
Disapproved	12,361	17,501	0.413937	=12,361/29,862
Totals	206,908	31,667	0.111189	=26,527/238,575
2016				
Approved	201,679	13,258	0.061683	=13,258/214,937
Disapproved	10,933	21,427	0.337855	=10,933/32,360
Totals	212,612	34,685	0.097822	=24,191/247,297
2015 and 2016				
Approved	383,958	39,692	0.093691	=39,692/423,650
Disapproved	30,002	32,220	0.482177	=30,002/62,222
Totals	413,960	71,912	0.143441	=69,694/485,872

major areas for the application of educational data mining: for improving student models; improving the domain model; studying pedagogical support using learning software; and for 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, Vellido, Nebot, and Mugica (2007) suggest the following tasks/applications for educational data mining: to evaluate student academic performance; to provide complementary courses according to the student's learning behavior; to evaluate educational resources available on Web courses; to give feedback to teachers and students in distance education courses; and to address atypical behaviors of student learning.

Romero and Ventura (2010) 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 a Concept Map, Planning & Scheduling and Constructing Courseware. The most commonly used tasks are regression, classification, clustering, and association rules. The algorithms most used on data mining are decision trees, neural networks and Bayesian networks.

Fonseca and Namen (2016) identify factors that relate profiles and their influences - positively and negatively - on students' Mathematics learning. They used the KDD methodology, with a focus on the Data Mining stage. Some discovered patterns are analyzed and discussed in the conclusion of this paper.

Dutt, Ismail, and Herawan (2017) provide a systematic literature review on clustering algorithm and its applicability and usability in the context of EDM. Future insights are outlined based on the reviewed literature, and possibilities for further research are identified.

Slater, Joksimović, Kovanovic, Baker, and Gasevic (2017) provide a discussion of the importance of familiarizing oneself with multiple tools – a data analysis toolbox for the practice of EDM/LA research.

Asif, Merceron, Ali, and Haider (2017) use data mining methods to study the performance of undergraduate students, focusing on two aspects of their performance. First, predicting students' academic achievement at the end of a four-year study program. Secondly, studying typical progressions and combining them with prediction results. Two important groups of students have been identified: low achieving and high achieving. The results indicate that by focusing on a small number of courses that are indicators of particularly good or poor performance, it is possible to provide timely warnings and give support to low achieving students, and advice and opportunities to high performing students.

In this paper, we carry out a prediction analysis of the academic performance of High School Juniors (students in their 3rd year of high school), similar to the study conducted by Fernandes, Carvalho, Holanda, and Van~Erven (2017). In contrast to this study, the present study adds to the model generated in the Fernandes study, generating a new model with data referring to the 2016 school year. Furthermore, in order to minimize the possibility of overfitting, another model was generated with data from both years. The generation of these models will serve as a basis for policy making and pedagogical interventions to improve the performance of these students by the end of each school year, potentially reducing the number of failures.

4. Proposed methodology

The methodology used in this research is based on the traditional Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000). We chose CRISP-DM for the applicability of its steps in a design based on data mining for predicting outcomes of student performance at the end of a one year school cycle. The six-step CRISP-DM model is illustrated in Fig. 1 and described below:

1. *Business Understanding.* This initial phase focused on understanding the project goals and requirements from a business perspective. In this phase, we used the iEducar database, which is composed of several attributes related to each student, such as bimonthly grades, courses taken, and absences. This database is used in the State Department of Education of the Federal District (SDEFD). After retrieving this dataset, we defined the research goal – to lower the failure rate of students in their third year of high school. Predicting student performance is a means to achieving this goal, as well as observing attributes obtained at the beginning of the school year. Notably, the attributes in this first phase did not include initial grades, as it occurred before the start of the school year and this information was not available. This strategy allowed us to identify the students who had the highest probability of failing at the end of the year. With this type of information, education specialists can concentrate extra effort on these particular students early in the year.
2. *Data Understanding.* The data understanding phase began with an initial data collection and proceeded with activities designed to acquaint users with them. We used the H2O Flow[3], 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.
   In this phase, we conducted an analysis covering 17 variables, namely: Educational District – designated by the SDEFD; city where the school is based; school name; class period; the presence of special needs students in a class; type of classroom; student's name; sex; age; government grant status; student's place of residence – city; student's place of residence – neighborhood; the student's special needs if any; school subjects; grades for the first two months; absences; and whether the student passed/failed the school year. This descriptive analysis generated the domain of each variable (how many records and which), the average, variances, mode, and standard deviation. The results of that statistical descriptive analysis are described in Section 5.
3. *Data Preparation.* This phase covered all activities to build the final dataset (data that was fed into the modeling tool) from the initial raw data.
   The variables of the dataset were prepared in order to generate the models used in the next phase. The goal of this research was to compare the prediction capability of student performance considering two different sets of variables. Thus, the key idea of *Data Preparation* was to build two datasets, in which the first one contains only personal, social and geographical information of students,

[3] H2o flow, https://www.h2o.ai/h2o/h2o-flow/, accessed: 2017, July.

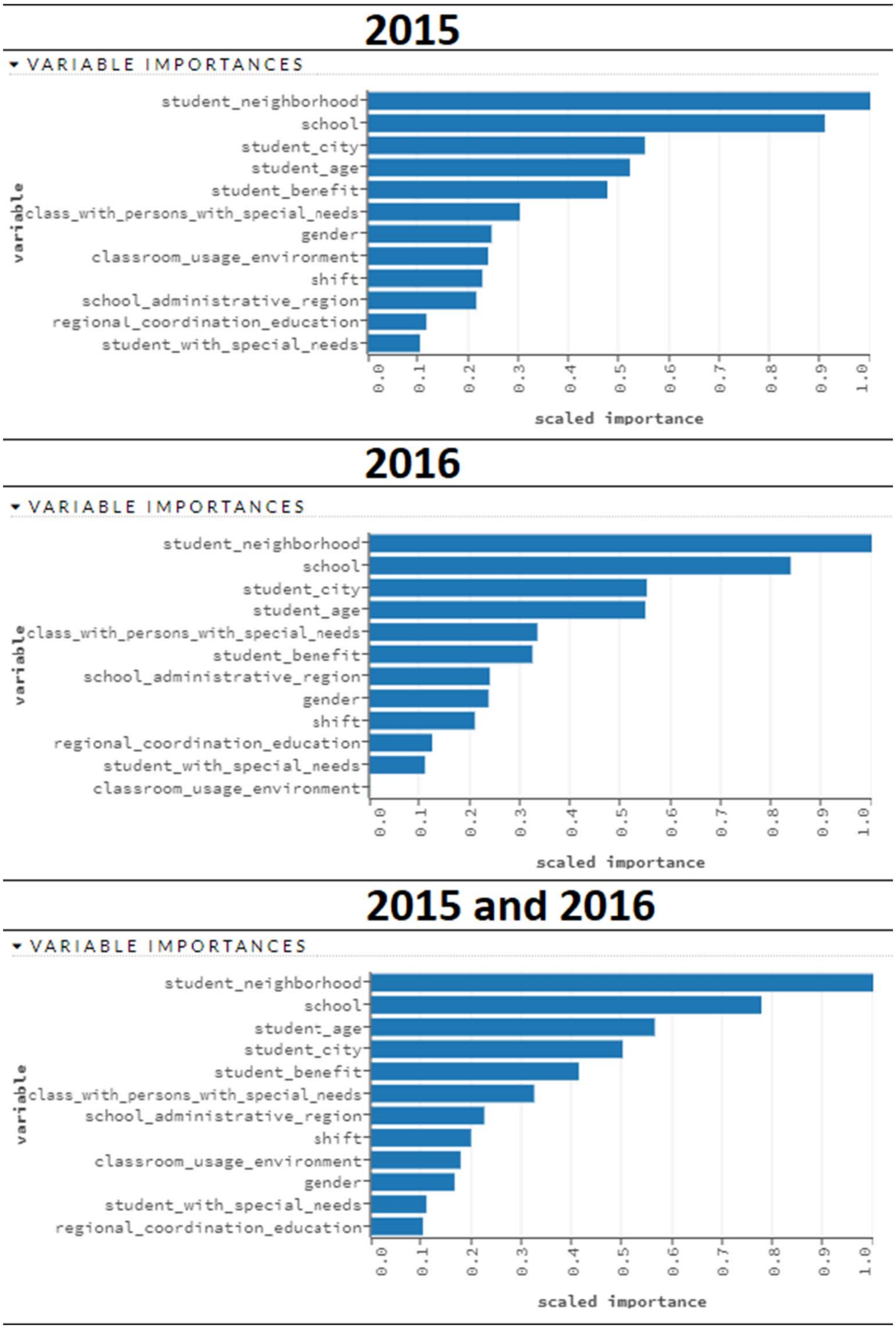

Fig. 3. Variable relevance for CM-I - beginning of the year.

while the other considers these variables, but includes academic variables, such as school subjects, grades and absences. A classification model was created for each dataset, enabling the evaluation and comparison of its performances for accurately predicting whether a student will pass/fail at the end of the school year. In other words, comparing the performance between two classification models allows us to verify the discrimination capability of the associated dataset's variables in relation to student performance. Details of the datasets obtained are provided in Section 5.

4. *Modeling*. In this phase, various modeling techniques were selected and applied, and their parameters calibrated to optimal values. In this paper, we chose the GBM[1] 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. The learning approach of a decision tree computes the relevance of each attribute of the dataset in relation to the students' ultimate evaluation: pass/fail. The implementation of the algorithm in H2O[4] was chosen for its ability to parallelize the classification process according to the machine's scores, making it possible to process a large amount of data and provide results efficiently. The accuracy of the classification

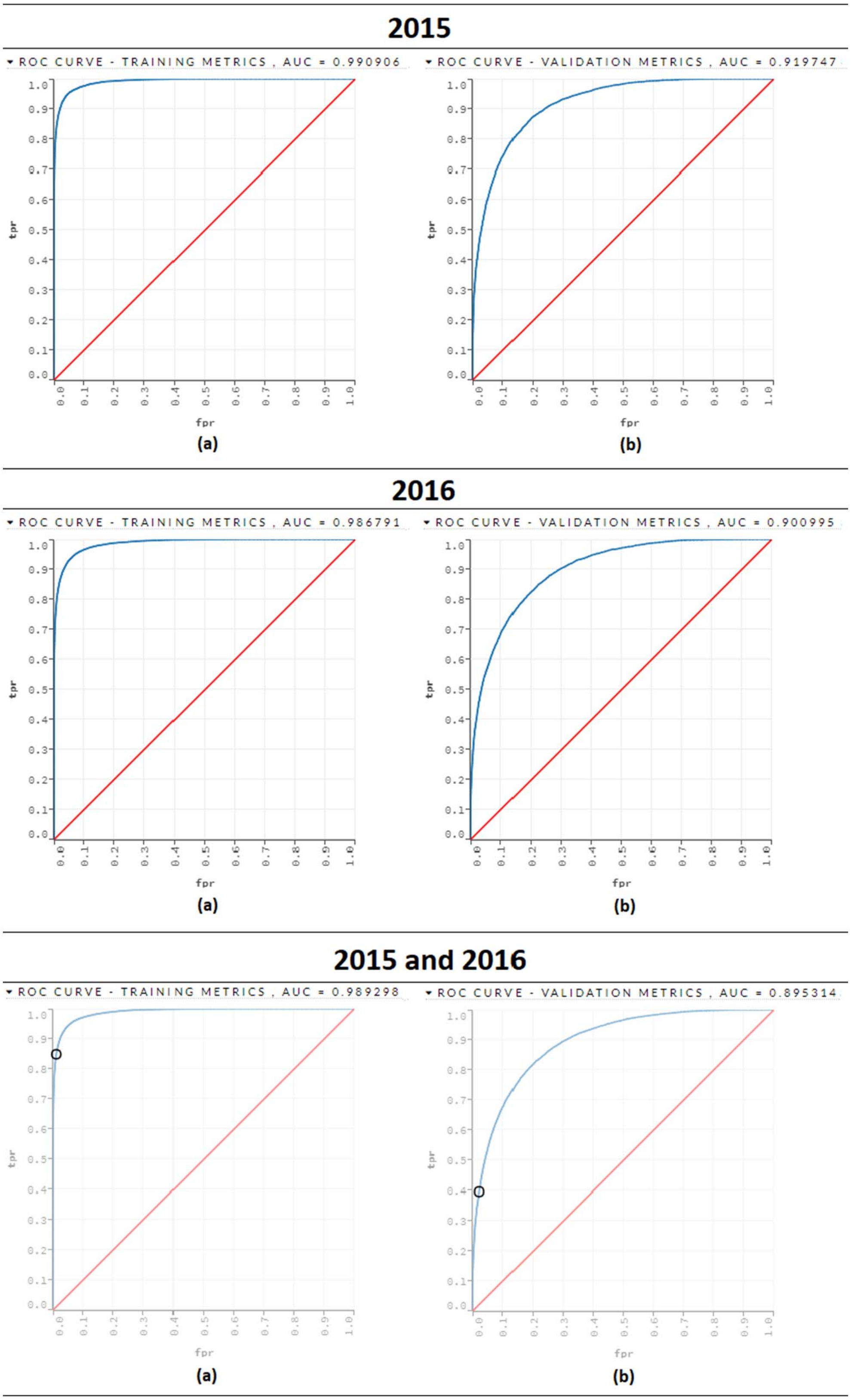

Fig. 4. ROC curve for CM-II - after 1 bimester - (a) training metrics and (b) validation metrics.

models were improved by means of the boosting technique, which is a flexible nonlinear regression procedure that helps to increase the accuracy of trees by sequentially applying weak classification algorithms to incrementally changing data. As a result, a series of decision trees were created, which resembled an ensemble of weak prediction models.

5. *Evaluation*. The key idea of this phase was to develop high quality models from the perspective of data analysis, and to evaluate if the generated model solved the problems raised in the Business Understanding phase. The generated classification models were

Table 3
Confusion matrix considering CM-II.

	Approved	Disapproved	Error	Rate
2015				
Approved	188,105	20,608	0.098738	=20,608/208,713
Disapproved	12,816	17,046	0.429174	=12,816/29,862
Totals	200,921	37,654	0.140099	=33,424/238,575
2016				
Approved	197,213	17,724	0.082461	=17,724/214,937
Disapproved	12,633	19,727	0.390389	=12,633/32,360
Totals	209,846	37,451	0.122755	=30,357/247,297
2015 and 2016				
Approved	396,499	27,151	0.064088	=27,151/423,650
Disapproved	24,924	37,298	0.400566	=24,924/62,222
Totals	421,423	64,449	0.107178	=52,075/485,872

evaluated using the Receiver Operating Characteristic (ROC) curve (Bradley, 1997). A discussion over the experimental results is provided in Section 5.

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.

The CRISP-DM is thorough and documented. All phases are organized, structured, and defined, so that the project may easily be understood and revised (Santos & Azevedo, 2005). For this study, all the CRISP-DM phases were used except the deployment phase, which will be mentioned in future works.

5. Results

In this section, we describe the results obtained in some specific phases of the proposed methodology: data understanding, data preparation and evaluation.

In the **Data Understanding** phase, as shown in Table 1, we obtained the quantitative information of data instances on nominal variables and the averages on numeric variables referring to the 17 variables of the data set for the years 2015 and 2016. We collected 238,575 records in the year 2015 and 247,297 records in the year 2016. There was an increase of 834 High School Juniors from 2015 to 2016. The failure rate increased from 12.5168% in 2015 to 13.0854% in 2016. Given that the purpose of this paper is to present findings that may help to reduce this failure rate, these data are essential for setting goals for the coming year.

In **Data Preparation**, we selected the variables for constituting two datasets, which are summarized below:

- Dataset I (DS-I): students are described by variables collected prior to the beginning of the school year.
- Dataset II (DS-II): students are described by the variables of DS-I and by new variables that became available 2 months after the beginning of the school year, such as school subjects, grades, and absences.

DS-I was used to train the classification model I (CM-I), which identifies the probability of student failure at the end of the year based on information prior to the start of the school year, as shown in the Table 1. Therefore, the “scholastic” variables, such as school subjects, grades and absences, were not available and so not included in DS-I. DS-II was used to train the classification model II (CM-II), which also determined the probability of a student failing at the end of the year. However, DS-II included the aforementioned “scholastic” variables, since it was used after the start of the school year and this information had become available. For both datasets, the students' names were not included, since they are useful only to identify a student, but not for identifying patterns. Furthermore, each dataset was generated for the years 2015 and 2016, and both years 2015–2016 together. This strategy allowed us to extract knowledge about student performance in three different periods over the two school year cycles.

In the **Evaluation** phase, we present some experimental results for assessing the performance of the classification models CM-I and CM-II for the years 2015, 2016, and for the period that spans them both, 2015–2016. Results using CM-I for 2015, presented a failure rate of 12.51%. Results from 2016 presented a failure rate of 13.08% and the average between both of the years was 12.80%, as seen in the data understanding phase. In addition, using CM-I, we found that in 2015 the training data – ROC curve – was 0.967168. In 2016 the curve measured 0.950594, while for 2015–2016 it measured 0.943076, as can be seen in Fig. 2 (a). With regard to the validation data, our ROC curve in 2015 was 0.962933. Finally, the classification model for CM-I with regard to 2016 showed a failure rate of 0.942826 and for 2015–2016 this rate was measured at 0.936658, as presented in Fig. 2 (b). These measurements show that there was a good rate in the sensitivity of the successes of CM-I, shown in the confusion matrix of 2015, 2016 and 2015–2016 according to Table 2, which shows the amount of records corresponds to correct hits per class. We can also check each variable's degree of importance to generate the data for 2015, 2016 and 2015–2016, as shown in Fig. 3. These data demonstrate that the variables ‘neighborhood’, ‘school’, ‘city’, and ‘age’ are relevant factors affecting the outcome of a student's academic performance and their ultimate degree of success.

After generating CM-II for 2015, 2016 and 2015–2016, we found that the training data ROC curve measured 0.990906, 0.986791 and 0.989298, respectively, shown in Fig. 4 (a), and the validation data ROC curve was 0.919747, 0.900995 and 0.895314, respectively, 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, respectively, in Table 3, 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, which presents that the variable ‘ grade’ resulted in a higher degree of importance than the variables ‘neighborhood’ and ‘school’, but the variables ‘school subject’ and ‘absence’ did not show a higher degree of importance than either of the variables ‘grade’ or ‘neighborhood’.

We can conclude that, with the generation of the classification model CM-I, consisting of variables obtained prior to the beginning of the school year, the most important variables for determining academic success are ‘neighborhood’, ‘school_subject’, ‘city’, ‘age’, demonstrating that a student's demographics – the social environment in which he or she circulates – directly influences the teaching-learning process.

On the other hand, the experimental results of the classification model CM-II showed that the variables ‘grade’, ‘absence’ and ‘school subject’, this latter differentiating DS-II from DS-I, gain greater expressiveness when we want to predict the final results of the students. However, the ‘absence’ and ‘school_subject’ variables do not show a higher degree of importance than the ‘neighborhood’ and ‘school’ variables. This is the case because, in the public schools of the Federal District, to pass, students need a grade average of 5 or more and 75% attendance.

6. Conclusion

This paper presented a methodology for analyzing the predictive performance of students in public schools of the Federal District of Brazil. The proposed methodology is based on the CRISP-DM and employed a dataset obtained from a repository of the State Department of Education of the Federal District of Brazil. Initially, a descriptive statistical analysis was carried out in order to obtain quantitative information of these data. To this end, we built two datasets: the first containing only attributes collected prior to the beginning of the school year; the second contained those same attributes, but included a few

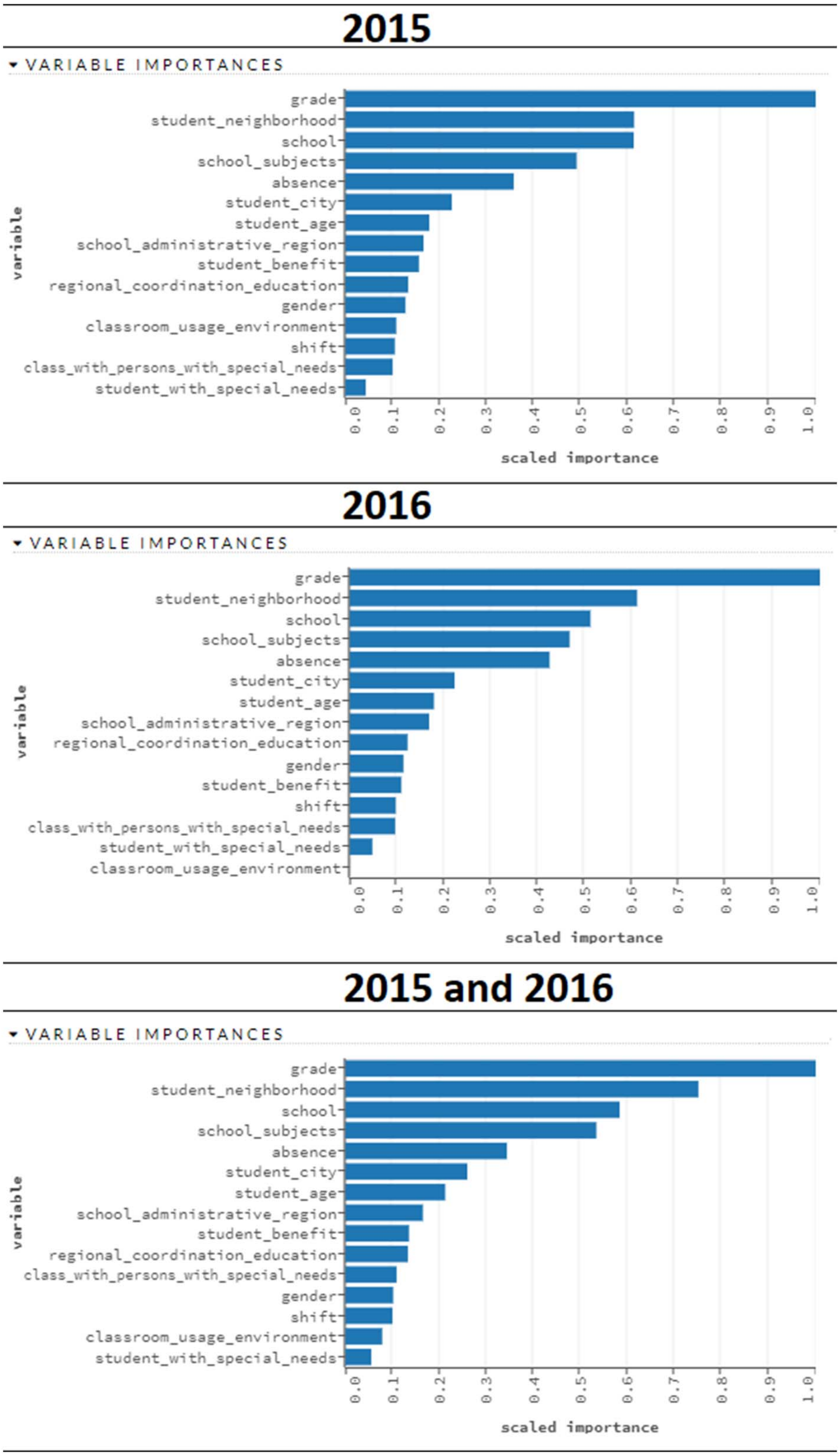

Fig. 5. Variable relevances obtained from CM-II, considering some variables obtained after 2 months to the beginning of the scholar year.

new variables, such as 'absences', ' grades' and 'school subjects'. In this case, these additional attributes were obtained 2 months after the beginning of the school year. Subsequently, we built a classification model based on the Gradient Boost Machine (GBM) for each dataset to predict student performance, allowing us to compare the predictive capability at two different moments in the academic cycle.

Experimental results indicated that the attributes identified prior to the beginning of the school year (first dataset) were relevant contributors to the failure rates. Particularly, the variables 'neighborhood' (student's residence) and 'school' were the main factors that affect the student's failure rate. This information can be useful for specialists interested in understanding that the student's ability to access a school or not, and the student's housing situation are relevant to the student's performance. Furthermore, the inclusion of academic variables to the social and personal attributes reveals that the student's grade is the most important variable for predicting their performance, followed by their attendance record. This can be explained due to the fact that these variables are direct and institutionalized requirements for the students to pass at the end of the year.

We anticipate that our research will provide meaningful information to guide coordinators, teachers and managers when making decisions concerning the school year, courses offered, public policies in education, etc. Moreover, the variables identified here as relevant factors in whether a student passes or fails at the end of the school year can improve the efficiency of pedagogical support for the student body. This added support can especially benefit students with learning difficulties, thus increasing their chances of passing at the end of the school year and reducing the failure rate in general.

Future work aims to use the proposed methodology within the State Department of Education of the Federal District. It is important to convey the knowledge obtained in this research to education specialists, so that they can make appropriate management decisions with regard to the public high schools. Furthermore, we intend to study other variables that can affect student performance, as well as other well-known data mining techniques. Finally, we plan to extend the application of the proposed methodology to an educational dataset of college undergraduate students.

References

Asif, R., Merceron, A., Ali, S., & Haider, N. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*(Supplement C), 177–194.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Baker, R. S. J. D. (2010). Data mining for education. *International encyclopedia of education, 7*(3), 112–118.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition, 30*(7), 1145–1159.

Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. *Evolution of teaching and learning paradigms in intelligent environment,* 183–221.

Chakraborty, B., Chakma, K., & Mukherjee, A. (2016). A density-based clustering algorithm and experiments on student dataset with noises using rough set theory. *IEEE International Conference on Engineering and Technology, 2016*, 431–436.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Crisp–dm 1.0 step-by-step data mining guide*.

Dutt, A., Ismail, M., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access, 5*, 15991–16005.

Fernandes, E., Carvalho, R., Holanda, M., & Van Erven, G. (2017). Educational data mining: Discovery standards of academic performance by students in public high schools in the federal district of brazil. *World Conference on Information Systems and Technologies,* 287–296.

Fonseca, S., & Namen, A. (2016). Data mining on inep databases: An initial analysis aiming to improve brazilian educational system. *Educação em Revista, 32*(1), 133–157.

Hanna, M. (2004). Data mining in the e-learning domain. *Campus-wide information systems, 21*(1), 29–34.

Hung, J.-L., & Crooks, S. M. (2009). Examining online learning patterns with data mining techniques in peer-moderated and teacher-moderated courses. *Journal of Educational Computing Research, 40*(2), 183–210.

Kabra, R., & Bichkar, R. (2011). Performance prediction of engineering students using decision trees. *International Journal of Computer Applications, 36*(11), 8–12.

Koedinger, K., Cunningham, K., Skogsholm, A., & Leber, B. (2008). An open repository and analysis tools for fine-grained, longitudinal learner data. *Educational Data Mining,* 157.

Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., & Riera, T. (2014). A system for knowledge discovery in e-learning environments within the European higher education area-Application to student data from open university of madrid, udima. *Computers & Education, 72*, 23–36.

Merceron, A., & Yacef, K. (2005). Educational data mining: A case study. *AIED,* 467–474.

Portal do software público brasileiro – i-educar. Retrieved from https://softwarepublico.gov.br/social/i-educar (Accessed: 2017, July)

Raghavan, S. N. (2005). Data mining in e-commerce: A survey. *Sadhana, 30*(2 & 3), 275–289.

Reis, E., Melo, P., Andrade, R., & Calapez, T. (1999). *Estatística aplicada*. Sílabo.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601–618.

Romero, C., Ventura, S., & De Bra, P. (2004). Knowledge discovery with genetic programming for providing feedback to courseware authors. *User Modeling and User-Adapted Interaction, 14*(5), 425–464.

Santos, M. F., & Azevedo, C. S. (2005). *Preâmbulo [a]" data mining: descoberta de conhecimento em bases de dados".* FCA editores.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics, 42*(1), 85–106.