Using Data Mining Techniques to Perform School Dropout Prediction: A Case Study

28

Renato Carauta Ribeiro and Edna Dias Canedo

Abstract

School Dropout is a severe problem for educational institutions. Institutions need to be able to measure and reduce dropout rates. Currently, annual expenses with dropout reach R$ 415 million in Brazilian currency. The purpose of this article is to identify the factors that affect students who drop out of the University of Brasilia (UnB) and Machine Learning to provide a model for predicting which students will drop out of undergraduate courses. With this, actions can be taken to reduce the dropout rate. The result of this work demonstrates that the courses with the most credits (workload), longer time to complete (5–6 year courses) and student's poorer academic performance (poor grades) influences student dropout rate. Also, social factors, such as quota holders or non-quota holders, also influence the dropout rate of undergraduate students at the University of Brasília (UnB).

Keywords

Educational data mining · Academic performance · Apriori · CRISP-DM · Machine learning

R. C. Ribeiro (✉)
Computer Center, University of Brasília (UnB), Brasília, DF, Brazil
e-mail: rcarauta@unb.br

E. D. Canedo (✉)
Computer Science Department, University of Brasília (UnB), Brasília, DF, Brazil
e-mail: ednacanedo@unb.br

28.1 Introduction

Dropout is an increasingly complex social problem for education professionals. You need to know why some students are unable to complete their studies. Several factors can influence school failure, such as: economic, social, family, educational condition, psychological profile, among others. For this reason it is a problem that is difficult to solve [2]. According to Rumberger [29], dropping out is seen as a socio-educational problem, and most people who drop out of school limit their economic well-being and lifelong social growth severely. The consequences of this can lead to billionaire costs for governments.

In Brazil, dropout is a serious problem, not only social, but also financial. Due to the containment of spending and budgetary constraints that the current government has made on education, the investments made in the educational area must have the desired effect. Currently, fewer undergraduate students complete the course. Expenditure generated by the dropout rate in federal educational institutions is around R$ 415 million [24]. Brazilians agree that the Public Federal Brazilian Universities high dropout rates require urgent solutions as do the appallingly low levels of work readiness for a large number of people. Even educators, educational managers and policy makers agree that the Educational System is in desperate need of reform [20].

Several studies were carried out in the area. Among them, one of the main ones that seeks a solution to this problem, focusing on the causes and possible interventions, is the path analysis model proposed by Tinto [32]. The model suggests that student social and academic integration in the educational institution is one of the main factors that determine the success and completion of the course. In the educational context, Data Mining (DM) is called Educational Data Mining (EDM). EDM is concerned with developing methods for exploring data in an educational context using

S. Latifi (ed.), *17th International Conference on Information Technology–New Generations (ITNG 2020)*, Advances in Intelligent Systems and Computing 1134,
https://doi.org/10.1007/978-3-030-43020-7_28

methods to understand better students and the settings in which they learn. EDM techniques can be used to create predictive models [26].

Due to the complexity of analyzing the many factors that lead to school dropout, the use of the Machine Learning technique is one of the most effective ways to obtain the desired results in containing this condition. Machine Learning serves as a fundamental tool for information extraction, data pattern recognition, and prediction [15]. Several classification algorithms provide a better level of accuracy (neural networks, SVM, k-means, and others.). All of these algorithms are black-boxed, meaning they hide algorithm details but provide excellent accuracy and facilitate the implementation of predictive models [11].

Annually at the University of Brasília (UnB) an average of 12,600 students enter. In 2017 UnB had a total of 53,657 students [8]. The UnB Teaching and Undergraduate Degree (DEG) Has developed some studies to verify the dropout of students in some undergraduate courses. In these studies, were evaluated students who entered from 2002 to 2008. The survey concluded that according to the course, students dropout rate is over 50% [7]. The objective of this work is to verify the amount of school dropout in the undergraduate courses of UnB. This paper presents a future forecast of the percentage of students who not finish their undergraduate degree at UnB. This analysis enables UnB to draw up a plan for improving its student retention policy and provide an overview of the courses that have the highest dropout rates and lead to the undergraduate degree (DEG) discussion on how to improve our performance and motivate our students to complete their courses.

28.2 Background and Related Works

Data Mining (DM) is the analysis of a dataset to find relationships and summarize data in new ways that are understandable and bring business benefits. Data mining is called a "secondary" data analysis because it deals with data that has already collected. Usually, no new data is created [17]. The main goal of DM is to discover relevant patterns and knowledge from a large amount of data. Data sources can range from structured data stored in databases to unstructured data [16].

To make predictions and discover patterns through DM several tasks can be classified into two categories: descriptive and predictive. Descriptive DM tasks characterize data properties in a target data set. Predictive DM tasks use induction in current data to make predictions [16]. Data mining, usually defined in the broader context of Knowledge in a Databases (Knowledge Disco-very in Databases-KDD). KDD is a multi-step process [17]: (1) **Data selection**: The data needed to solve the problem to be solved by the DM is selected. (2) **Data preprocessing**: The data obtained must first be pre-processed to transform it into an appropriate format for mining. Some of the main preprocessing tasks are: cleanup, attribute selection, attribute transformation, data integration, and others. (3) **Data Mining (Pattern Extraction)**: This is the intermediate step that identifies the entire process. During this step, Are applied data mining techniques to pre-processed data. (4) **Post-processing**: This is the final step in which the results obtained or model are interpreted and used to make decisions relevant to the business area.

An essential step in the DM process is the search for the relationship between the attributes to have useful representations of some aspects of the data. This process involves some steps [17]: (1) Determine the nature and structure of the representation to be used; (2) Decide how to quantify and compare how different representations fit the data; (3) Choose an algorithmic process to optimize the scoring function; (4) Decide what data management principles are needed to implement the algorithm efficiently.

Educational Data Mining (EDM) is an application of DM techniques for educational problems. The goal is to solve problems and challenges in the area of education. EDM has been a large area of research were several areas of knowledge that seek to analyze the large volume of educational data to solve education problems [27]. The EDM process transforms raw data into potentially relevant data for analysis of educational research involving DM. The steps for data analysis are similar to other areas. Are done preprocessing, data mining, and post-processing. EDM uses the same rules as DM, such as association rules, text mining, and more. Besides, EDM has been making discoveries with modeling and integration of structured modeling psychometric variables. These techniques are uncommon in DM [27].

EDM aims to improve the learning process and gain a deeper understanding of educational phenomena. These phenomena are difficult to quantify because there are a multitude of different types of data available [27]. EDM is an emerging discipline that is increasingly developing methods for exploiting large-scale unique data from the educational context, using methods to understand better students and the settings in which they learn [27]. Today there is a wide variety of educational systems and environments: classrooms, LMS e-learning, online education systems. It is also available a significant content of online learning such as quizzes, forums, virtual environments, among others, which are increasingly used in the educational context. This wealth of tools and content makes more and more information available for EDM to review [27].

Machine Learning is the technique of teaching the machine to learn—by changing its structure—in programs or data so that its expected future is an improvement in perfor-

mance. Such changes involve recognition, diagnosis, planning, robot control, predictions, among others [23]. The machine learning technique is designed to make computers adapt to specific actions to improve accuracy, and the machine can learn by itself the pattern taught to it [21]. To be effective, active machine learning involves several different disciplines, such as [23]: (1) *Statistics:* This is the discipline that selects the samples that should be used for machine learning. (2) *Mental Models:* This is the area that studies how closely machine modelers approach the way living brains learn. (3) *Adaptive Control Theory:* It is the discipline that studies the problem of controlling a process with unknown parameters that must be estimated during the operation. (4) *Psychological Models:* Study the performance of humans in various learning tasks. (5) *Artificial Intelligence:* It is the discipline that is concerned with machine learning. (6) *Evolutionary Models:* It is the discipline that studies techniques that model certain aspects of biological evolution and applies this to machines to improve the performance of computer programs.

28.2.1 Related Works

Márques et al. [22] compared the algorithms for data mining using a new approach called ICMR2. The focus of the work is to verify the causes of dropout for university students. The purpose of the ICMR2 approach is to determine which students are more likely to drop out and why. The developed methodology has the purpose of improving the prediction of possible school dropouts of students. The results of this study show that the algorithm created was able to predict the dropout of students from four to 6 weeks. It is reliable enough to be used with production data. Breiman [4] presents in his paper the definition and use of the Random Forest algorithm, one of the most commonly used algorithms in Data Mining. The Random Forest is defined as a classifier where each tree casts a vote for the most popular class among database attributes. Random Forest is a useful forecasting tool.

Archambault et al. [3] presented a case study using samples from French Canadian students to assess, through statistical analysis, the engagement of these students and prospects of dropping out. A multidimensional approach is used to analyze the study, which analyzes the student through multiple attributes [3]. This study was able to predict the dropout rate reliably. Of the three specific dimensions, only behavioral engagement made a significant contribution to the prediction equation. The study also concludes the robustness in multidimensional quantitative attributes for student prediction. According to Cornell et al. [6], one of the factors affecting dropout, especially at the elementary and middle levels, is bullying. This study demonstrates how bullying can affect student permanence in school, hamper their academic growth, and further professional development. The study concludes that high school bullying is one of the major factors that negatively affect student performance and is the main variable by which students drop out of school.

Ge et al. [15] provided a review of data mining and analytical applications in the industry in recent decades. Are explored eight unsupervised algorithms and ten supervised algorithms. Several perspectives are highlighted and discussed about the analyzed algorithms. It has been found that in addition to supervised and unsupervised approaches, semi-supervised machine analysis has recently been introduced, which become more popular soon. According to Romero and Ventura [28], Educational Data Mining (EDM) is concerned with developing methods for exploiting unique types of data from the educational environment. EDM differs from DM in using specific techniques for analyzing educational data. The article provides an overview of EDM, along with applications, tools, and future perspectives. The work by Shahiri et al. [30] aims to systematically review the literature on student performance prediction using data mining techniques to predict student performance better. This article focuses on how to select attributes for educational data mining. The article demonstrates that most researchers use cumulative grade point average (CGPA) and internal assessment as the data set for forecasting. The most used method in EMD is the classification method. The most commonly used classification techniques are: Neural Networks and Decision Trees. These two techniques are often used to predict student performance. The work presented by Fernandes et al. [14] presents a predictive analysis of the academic performance of public school students in the Federal District of Brazil during the 2015 and 2016 school periods. Data were collected from each school year for the analysis. The model used was the Gradient Boosting Machine (GBM). The result of this research showed that while attributes such as 'class' and 'absences' were the most relevant for year-end forecasting, the academic results of demographic attribute performance reveal that 'neighborhood', 'school' and 'age' are also potential indicators of a student's academic success or failure.

The result of this work differs from previous work in that it analyzes a specific context in which it analyzes a sample that contains data from students of undergraduate courses at UnB. This sample contains the negative student grades (below the passing grade), the university entry form, whether or not the student is a quota holder, and the number of course credits. These variables were relevant in the analyses performed in this work to verify the possibility of the student completing his undergraduate course or leaving it before its conclusion. It is possible to state that students with grades below the average required to pass a course, quota holders and who have taken courses with a higher amount of credits, have a

greater tendency to drop out of their courses. In this paper, we analyzed several models for problem-solving, and the best model for predicting school dropout was the GBM model, also used in production developed by Fernandes et al. [14].

28.2.2 Method

This paper presents a case study carried out in the context of the University of Brasília. References for this study were searched using the Web of Science database. The researched papers cover the period from 2009 to 2019 [13]. (1) "data mining and school dropout"—6 results; (2) "school dropout"—17 results; (3) "data mining"—26.047 results; (4) "school dropout" psychology area—191 results. The integrated model and evidence validation were used to integrate the main papers addressed in the researched areas. Were used the coupling approach and the co-citation approach. The software used to form the research networks was VOSviewer 1.6.11, which reads the database. In this paper, we used the database Web of Science. We generated a bibliometric map containing the primary references of each of the searched topics [33].

28.3 Proposed Methodology

The Cross-Industry Standard Process for Data Mining (CRISP-DM) reference model provides an overview of the life cycle of a DM project. It consists of several phases that provide a faster, more reliable data mining process with greater management control [5]. The CRISP-DM model contains the phases of a project, their respective tasks, and the relationship between them. The life cycle of a data mining project consists of six phases [5]: (1) **Business Understanding:** This early phase focuses on understanding project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition and preliminary plan designed to achieve the goal. (2) **Understanding the data:** It begins with an initial data collection and proceeds with activities to familiarize it with the data, identify the data quality problem, discover first data ideas, or detect interesting subsets to form hypotheses for hidden information. (3) **Data Preparation:** Covers all activities to build the final data set from the initial raw data. Data preparation tasks can be performed multiple times. Tasks include configurable selection, attribute registration, transformation, and data cleansing. (4) **Modeling:** Modeling techniques are selected and applied and parameters are calibrated for optimal values. You must return to the data preparation phase until you reach a suitable model. (5) **Evaluation:** At this stage, the appropriate model is built, but before it is put into production it is necessary to evaluate it in more detail to ensure that it adequately meets business objectives. (6) **Deployment:** Consolidates the knowledge discovered with the model to be created. The goal is to consolidate knowledge about the data and present it in a way that can be useful to the business. It is essential to develop a model monitoring and maintenance plan to prevent misuse of mining data and to keep the model always up to date.

Initially, the understanding of the business be presented. That is, all the contextualization about school dropout focused on UnB, what problem it generates, and how the use of Data Mining helps the process to minimize the level of school dropout. Later the analysis and understanding of the data will be made. The data used will be data from students and undergraduate subjects, which will be useful for subsequent verification of the percentage of dropout in UnB courses. After defining and understanding the data to be used, the attributes should be selected and the necessary transformations made to use this data in the model that will be created. After data preparation, model design, training, and testing are required to verify that the model has acceptable levels of accuracy. In this step a training mass, an evaluation mass and a test mass are selected. Having a satisfactory result, this model is put into production for future predictions, otherwise a better understanding and selection of data mass attributes is required. The predictive model is created using some classification algorithms. Some algorithms with a supervised approach be selected. The efficiency of each algorithm be shown, and a comparison is made between each model created.

28.4 Results

28.4.1 Data Analysis

The Integrated Graduation System (SIGRA) [9] is the current UnB system that has data on undergraduate students, from mention, subjects to school history. This system uses the BD-Siac database, which is stored on a Data Base Management System (DBMS) SQL Server [31]. The data used were taken from this base through the SQL language [25]. The tables used in this database are presented in Table 28.1.

Table 28.1 Database tables used

Tables	Description
TB_Aluno	Table with student's personal data
HEQuadroResumoGra	Table that keeps the history of the students
Opcao	Table that holds the course options
DadosOpcao	Student option data
Curso	Table that holds the courses of UnB

The data taken from these tables aims to create a sample of students who did not complete the course and those who graduated. The most relevant data for school dropout prediction are the ones related to the student's school history and social condition. The total attributes found with the SQL query against the tables returned 156 attributes. The identifiers (ID) and foreign keys for each table have been removed. All attributes about date except the year of birth were taken from the mass of data that generate the model. Attributes that identify a student such as: academic record, parent's name, address, telephone, among others, were also removed. A descriptive analysis was performed with Pearson's correlation coefficient, and the attributes that had a high correlation with each other were removed to avoid the model generating overfit. The attribute *DadForSaidOpc* was defined, which represents the classifying variable that shows graduated students and those who left the undergraduate course. With the remaining attributes, a relevance analysis was performed between all attributes and the classifier attribute. The selected attributes are explained below:

- *aluforingunb* Shows student entry form;
- *alupne* Shows students who are or are not handicapped;
- *aludtnasc* The date of birth of the students;
- *alucotid* Checks which students are or are not quota holders;
- *alumunicipio* Shows the students municipality;
- *alupassaporte* Checks whether or not the student has a passport;
- *OpcCredFormat* Amount of credits to graduate;
- *OpcMinPerm* Minimum period of stay in the course;
- *OpcMaxPerm* Maximum length of stay in the course;
- *SumTR* Number of times the student has locked a discipline;
- *SumTJ* Number of times student justifiably locked course;
- *SumSR* Number of occurrences the student obtained SR;
- *SumII* Number of occurrences the student obtained II;
- *SumMI* Number of occurrences the student obtained MI;
- *SumDP* Number of occurrences in which the student was dismissed from any discipline;
- *SumCC* Number of times the student has earned credits;
- *SumAP* Number of Student Approvals;
- *SumSC* Selective subjects courses;
- *Duracao* Duration of each subject;
- *Turno* Course shift (morning or evening);

Figures 28.1, 28.2, and 28.3 show the graphs showing the percentage of students graduated and not graduated according to the amount of grades. The terms used in this study are classified as: No Performance (SR), Lower (II) and Lower Average (MI) [10]. It can be seen that the higher the number of negative grades, the greater the chance of dropping out of the course.

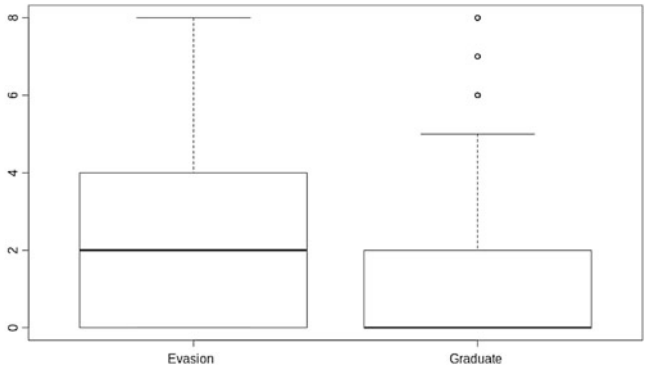

Fig. 28.1 Dropout and graduated with quantitative notes SR

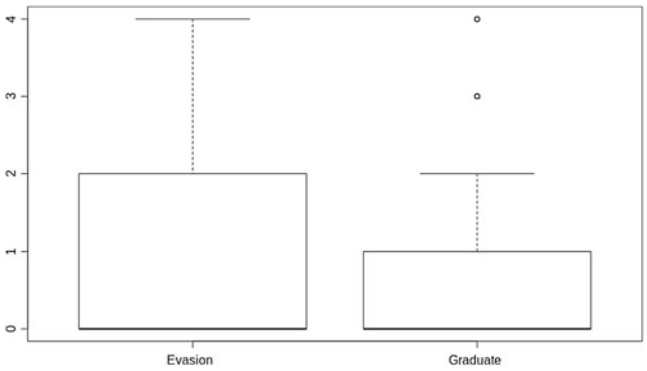

Fig. 28.2 Dropout and graduated with quantitative notes II

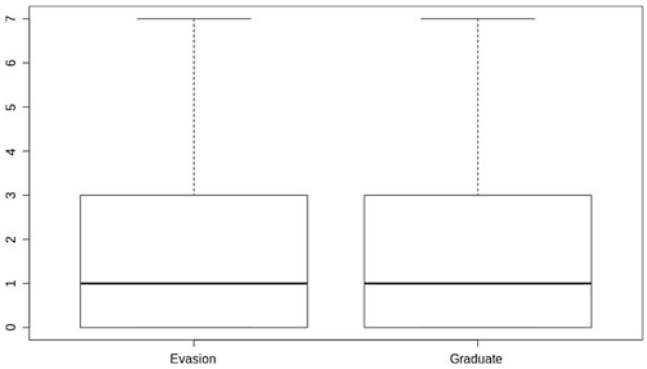

Fig. 28.3 Dropout and graduated with quantitative notes MI

28.4.2 Data Preparation

From all the selected attributes, those that most contribute to verifying the student's chance to drop out of the undergraduate course were selected. Attributes that have null values that are numeric have been replaced by "0", and those that are alphanumeric have been replaced by the character "A". It can be noted that the attributes such as student grades, quota holder or not, form of admission, foreign student or not, student social status, among other attributes, define the abandonment or retention of the student in undergraduate (Fig. 28.4).

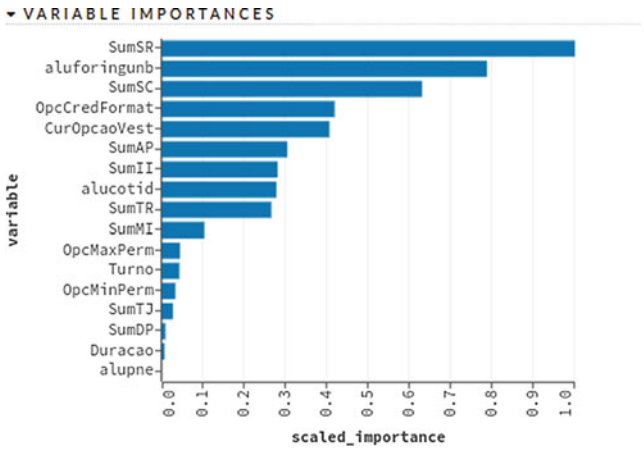

Fig. 28.4 Variable importances

Confusion Matrix (vertical: actual; across: predicted)

	Evasion	Graduate	Error	Rate
Evasion	2486	2458	0.497168	=2458/4944
Graduate	674	5399	0.110983	=674/6073
Totals	3160	7857	0.284288	=3132/11017

Fig. 28.5 Validation performance data resume from GLM model [12]

Confusion Matrix (vertical: actual; across: predicted)

	Evasion	Graduate	Error	Rate
Evasion	3174	1770	0.358010	=1770/4944
Graduate	653	5420	0.107525	=653/6073
Totals	3827	7190	0.219933	=2423/11017

Fig. 28.6 Validation performance data resume from GBM model [12]

28.4.3 Model and Evaluation

The model used was a model belonging to the supervised classification. We used the H2O package [1] with some algorithms to predict future school dropouts. (1) GLM: Generalized Linear Model; (2) GBM: Gradient Boosting Machine; (3) SVM: Support Vector Machines; (4) RF: Random Forest. For the generation of the model, we used a data sample with a total of 35,646 students, who graduated or dropped out of UnB between 2006 and 2018. we separated this sample into three parts: 50% for training, 30% for validation, and 20% for testing. Data were applied to the GLM, GBM, SVM, and RF models [12]. Each of these models generated a specific confusion table with the hit and miss rates, as shown in Figs. 28.5, 28.6, 28.7, and 28.8.

Of the four models analyzed, the models that had the highest accuracy were the GBM model and the RF model. Due to the lower number of errors of graduated and abandoned students, the GBM model was chosen to deploy. The ROC curve of the GBM model had an accuracy of 86% with the attributes chosen for its construction, as shown in Fig. 28.9.

Confusion Matrix (vertical: actual; across: predicted)

	Evasion	Graduate	Error	Rate
Evasion	2594	2350	0.475324	=2350/4944
Graduate	723	5350	0.119052	=723/6073
Totals	3317	7700	0.278933	=3073/11017

Fig. 28.7 Validation performance data resume from SVM model [12]

Confusion Matrix (vertical: actual; across: predicted)

	Evasion	Graduate	Error	Rate
Evasion	3071	1873	0.378843	=1873/4944
Graduate	597	5476	0.098304	=597/6073
Totals	3668	7349	0.224199	=2470/11017

Fig. 28.8 Validation performance data resume from random forest model [12]

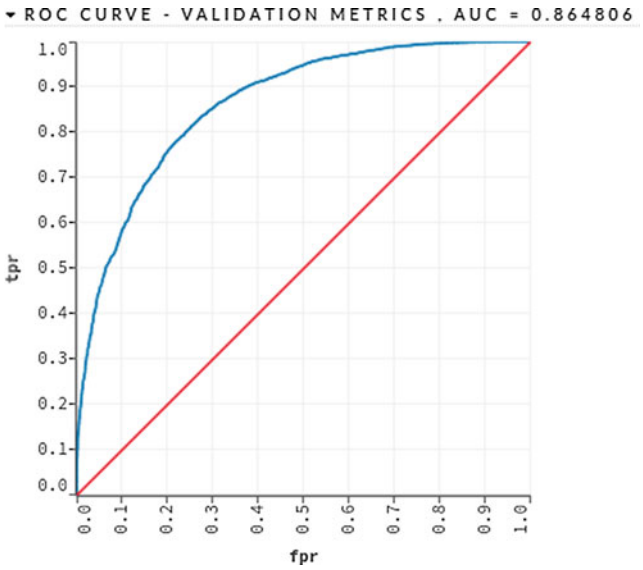

Fig. 28.9 ROC curve for validation metrics [19]

Applying the GBM model to undergraduate students at UnB from 2017/1, the dropout rate of undergraduate courses was around 54%. The output files were created in csv format by RStudio [18] with prediction and evasion probabilities. The prediction made by GBM shows that above 50%, the algorithm predicts that an undergraduate student will drop out of University. This work shows that measures to contain the dropout rate at UnB are necessary; for example, a more detailed monitoring of students who are more likely to drop out.

28.5 Conclusion

The objective of this study was to predict the dropout of undergraduate students from UnB. From the results achieved with the GBM model, it was possible to predict, with an acceptable confidence rate, whether the student will evade

or graduate from the undergraduate course. The CRIPS-DM model used in this paper collected data from the SIGRA system databases. The data were understood and prepared. After the creation of four models, one of them was chosen as the best, and a prediction was made with the students currently enrolled in UnB undergraduate courses. The GBM model predicted a 54% chance of students dropping out before completing UnB undergraduate courses. According to the generated model, factors such as the amount of course credits, the minimum and maximum amount of time required to complete an undergraduate degree, the student's entry form at UnB, the sum of negative grades obtained in the first Course subjects are some relevant factors that lead the student not to complete an undergraduate degree. Besides, it has been shown that social factors contribute to undergraduate dropout.

It was concluded that factors such as student's academic performance and the degree of difficulty of the undergraduate course are still the main factors of dropout, but social issues are also relevance for the completion or dropout of an undergraduate course. Based on this forecast it is possible, in future works, a more in-depth analysis for each undergraduate course aiming at obtaining tools to take preventive measures with the objective of minimizing the dropout rate at UnB or other higher education institution.

References

1. Aiello, S., Eckstrand, E., Fu, A., Landry, M., Aboyoun, P.: Machine learning with R and H2O. H2O booklet (2016)
2. Aloise-Young, P.A., Chavez, E.L.: Not all school dropouts are the same: ethnic differences in the relation between reason for leaving school and adolescent substance use. Psychol. Schools **39**(5), 539–547 (2002)
3. Archambault, I., Janosz, M., Fallu, J.-S., Pagani, L.: Student engagement and its relationship with early high school dropout. J. Adolesc. **32**, 651–670 (2009)
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
5. Chapman, P., Clinton, J.M., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C.R.H., Wirth, R.L.: CRISP-DM 1.0: step-by-step data mining guide (2000)
6. Cornell, D., Gregory, A., Huang, F., Fan, X.: Perceived prevalence of teasing and bullying predicts high school dropout rates. J. Educ. Psychol. **105**, 138 (2013)
7. CPA, U.: Pesquisa de retenccão e evasão (2018)
8. de Brasília (unB), U.: Anuário estatístico 2018: um raio-x da unb (2018)
9. de Brasília, U.: Sistema de graduação (SIGRA)
10. de Brasília Introdução a Comunicação, U.: Avaliação
11. Delibašić, B., Vukićević, M., Jovanović, M., and Suknović, M.: White-box or black-box decision tree algorithms: which to use in education? IEEE Trans. Educ. **56**(3), 287–291 (2013)
12. Ellis, N., Davy, R., Troccoli, A.: Predicting wind power variability events using different statistical methods driven by regional atmospheric model output. Wind Energy **18**(9), 1611–1628 (2015)
13. Felizardo, K.R., Nakagawa, E.Y., Fabbri, S.C.P.F., Ferrari, F.C.: Revisão Sistemática da Literatura em Engenharia de Software: Teoria e Prática. Elsevier, Brasil (2017)
14. Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., and Erven, G.V.: Educational data mining: predictive analysis of academic performance of public school students in the capital of Brazil. J. Bus. Res. **94**(C), 335–343 (2019)
15. Ge, Z., Song, Z., Ding, S.X., Huang, B.: Data mining and analytics in the process industry: the role of machine learning. IEEE Access **5**, 20590–20616 (2017)
16. Han, J., Kamber, M., Pei, J.: Data Mining Concepts and Techniques, 3rd edn. Elsevier, Amsterdam (2012)
17. Hand, D.J.: Data Mining Based in Part on the Article "Data mining" by David Hand, Which Appeared in the Encyclopedia of Environmetrics. American Cancer Society, New York (2013)
18. Horton, N.J., Kleinman, K.: Using R and RStudio for Data Management, Statistical Analysis, and Graphics. Chapman and Hall/CRC (2015)
19. LeDell, E., Petersen, M., van der Laan, M.: Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. Electron. J. Stat. **9**(1), 1583 (2015)
20. Manhães, L.M.B., da Cruz, S.M.S., Zimbrão, G.: The impact of high dropout rates in a large public Brazilian university—a quantitative approach using educational data mining. In: CSEDU, vol. 3, pp. 124–129. SciTePress, Setúbal (2014)
21. Marsland, S.: Machine Learning: An Algorithmic Perspective, 2nd edn. Chapman & Hall/CRC
22. Márquez, C., Cano, A., Romero, C., Mohammad, A., Fardoun, H., and Ventura, S.: Early dropout prediction using data mining: a case study with high school students. Expert Syst. **33**, 107–124 (2016)
23. Nilsson, N.J.: Introduction to machine learning: An early draft of a proposed textbook, pp. 175–188. http://robotics.stanford.edu/people/nilsson/mlbook.html
24. Prestes, E. M. D. T., Fialho, M.G.D.: Evasão na educação superior e gestão institucional: o caso da universidade federal da paraíba. Ensaio: Avaliação e Políticas Públicas em Educação **26**(100), 869–889 (2018)
25. Rockoff, L.: The language of SQL. Addison-Wesley, Reading (2016)
26. 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)
27. 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)
28. Romero, C., Ventura, S.: Data mining in education. WIREs: Data Min. Knowl. Disc. **3**(1), 12–27 (2012)
29. Rumberger, R.W.: High school dropouts: A review of issues and evidence. Rev. Educ. Res. **57**(2), 101–121 (1987)
30. Shahiri, A.M., Husain, W., Rashid, N.A.: A review on predicting student's performance using data mining techniques. Proc. Comput. Sci. **72**, 414–422 (2015)
31. Tang, Z., Maclennan, J.: Data Mining with SQL Server 2005. Wiley, London (2005)
32. Tinto, V.: Leaving college: rethinking the causes and curse of students attrition. J. Adolesc. **2**, 269 (1987)
33. van Eck, N.J., Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics **84**(2), 523–538 (2010)