Educational Data Mining: A Profile Analysis of Brazilian Students

Edna Dias Canedo[1(✉)], Heloise Acco Tives Leão[2], Rhandy Rafhael de Carvalho[1], Ruyther Parente da Costa[1], Giovanni Almeida Santos[1], and Marcio Vinicius Okimoto[1]

[1] University of Brasília (UnB), Brasília, DF 70910-900, Brazil
{ednacanedo,giovannix}@unb.br, rhandyrafhael@gmail.com, ruyther@me.com, marciobtos@gmail.com
[2] Lutheran University Center of Palmas (CEULP), Palmas, TO, Brazil
heloise.acco@gmail.com

Abstract. This paper presents an analysis of data referring to the profile of Brazilian students in the year 2016, according to the Higher Education Census. The information provided by this census is used to carry out various analyses related to the current situation of Brazilian education, as well as the profile of students and institutions. In this work, we analyze the modality of courses offered by educational institutions, investigating the profile and the quantitative of students who have scholarship, and how they enter the courses. We also investigate the distribution of students according to the informed skin color/race. The method used to conduct this research involved the application of the steps proposed by the CRISP-DM Reference Model and the use of the Apriori association algorithm to identify and analyze the data set. The results show that 30% of students choose not to report their skin color; Of those who reported, 37.4% are white and 6% are black. Approximately 82% of the students entered higher education through the traditional method (entrance examination), and 82.3% of the male and 77.2% of the female opted for face-to-face courses.

Keywords: Educational Data Mining · Apriori algorithm · CRISP-DM · Brazilian students profile

1 Introduction

Several areas of knowledge generate data at an impressive pace. The large increase in data storage and collection has created a demand for techniques capable of extracting potentially useful information from the mass of data. Historically, the notion of mining databases to find useful data patterns has received many names, such as Knowledge Discovery in Databases (KDD), Knowledge Extraction, Data Harvesting, among others. Despite the different names, and in some cases different approaches, these terms have as main objective the extraction of knowledge from databases. Data mining is the non-trivial process of

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11619, pp. 473–488, 2019.
https://doi.org/10.1007/978-3-030-24289-3_35

identifying valid, potentially useful and understandable patterns in the mass of data, with a strong emphasis on working with large amounts of real data [20].

Data mining techniques are increasingly gaining significance in the education sector. As in many other sectors, higher education is discovering the potential impact of these techniques on the learning process and outcomes to move towards a university of the new era, providing educational policymakers with data-based models essential for supporting goals to enhance the efficiency and quality of teaching and learning [2].

Educational data mining (EDM) and Learning Analytics (LA) are two specific areas that are used to represent the use and the application of data mining in higher education and other educational settings. EDM can be defined as the application of traditional data mining techniques to the analysis of educational data, aiming at solving problems in the educational context [6]. Some applications of EDM include the development of e-learning systems [22], distance education, pedagogical support [15], educational data collection [10], student and teacher performance forecasts, evaluation of the quantity needed for the distribution of teachers versus number of schools, among others [9,18]. EDM techniques can reveal useful information to educators and government bodies to help them design or modify the curriculum structure of courses, schools and/or universities. They can help to identify at-risk students, identifying priority teaching and learning needs for different groups of students and teachers (teacher training), increasing the passing rates in schools and undergraduate courses, effectively evaluating institutional and teacher performance, maximizing the resources of schools and/or universities, and optimizing and renewing the curricular programs of the courses.

LA is a closely related endeavor, with somewhat more emphasis on simultaneously investigating automatically collected data along with human observation of the teaching and learning context. Overall, cyberlearning emphasizes the integration of learning sciences theories with these techniques in order to improve the design of learning systems and to better understand how people learn within them. The use of EDM and LA in the educational context has the potential to shape the existing models of teaching and learning by providing new solutions to the interaction problem. EDM and LA are used to offer more personalized, adaptive, and interactive educational environments to enhance learning outcomes, teaching and learning effectiveness, and optimizing institutional proficiency, as well as mapping both professor and student performance [25].

This paper investigates the profile of Brazilian students related to the teaching modality (face-to-face or distance learning) and the study period they chose to attend when they entered higher education. It also analyzes the profile of the students as to skin color, type of university entrance, and if they have a scholarship. To conduct this study, we defined the following research questions (RQ):

RQ.1: What is the profile of Brazilian higher education students related to the different teaching modalities offered (face-to-face or distance learning)?

RQ.2: What is the profile of Brazilian higher education students related to the existence of a scholarship?
RQ.3: What is the form of admission of students to educational institutions?
RQ.4: What is the distribution of Brazilian students according to skin color/race and gender?

This paper is organized as follows. Section 2 presents the state of the art related to educational data mining, as well as mining techniques. In addition, related works to this paper are presented. Section 3 details the method adopted to develop this work. Section 4 presents the results of the CRISP-DM phases. The conclusions and future work are presented in Sect. 5.

2 Educational Data Mining

Educational systems are increasingly engineered to capture and store data on users' interactions with a system. These data can be analyzed using statistical, machine learning, and data mining techniques. The development of computational tools for data analysis, standardization of data logging formats, and increased computation/processing power is enabling learning scientists to investigate research questions using this data. Educational Data Mining (EDM), and the application of Data Mining techniques in data from online education platforms or environments. EDM has emerged as a research area in recent years by researchers in several areas of knowledge seeking to analyze large volumes of data in order to solve educational research issues [30]. Overall, EDM draws on traditional statistical techniques and shares further challenges with other analytic uses of research data, such as: 1. Combining needed data from different systems, which can be difficult. 2. Achieving construct validity and interpretability of results. 3. Understanding consequential validity and use of results to drive decisions. 4. Deciding whether use of data to drive high-stakes and/or low-stakes decisions is warranted. 5. Establishing safeguards for privacy and ethics of data use [27].

EDM and the research area whose main focus is the development of methods to explore databases collected in educational environments, allowing a standard in the approach of educational data mining. Thus, it is possible to more effectively and adequately understand information about teachers, schools, management and students. It allows understanding how learners learn, the context in which learning occurs, how teachers are motivated and/or followed, and other factors that influence teaching learning. The EDM process converts raw data from educational systems into useful information that can be used by developers, teachers, educational researchers, etc. This process does not differ much from other areas of application to that of data mining because it is based on the same steps of the data mining process in general [30]. The educational data mining process comprises the following steps [26], [17]: **1. Preprocessing**: The data obtained with the educational environment must first be pre-processed to transform it into a format suitable for mining. Some of the main tasks of preprocessing are: cleaning, selection of attributes, transformation attributes, data

integration, and so on; **2. Data Mining (Extracting Patterns)**: It is the central step that identifies the whole process. During this stage, data mining techniques are applied to previously pre-processed data; **3. Post processing**: It is the final step in which the results obtained or model are interpreted and used to make decisions about the educational environment.

Data generated by students and teachers in e-learning environments can provide rapid and important insights into the performance, motivation and level of participation of students and teachers in a particular subject or course. These understandings may suggest significant changes and interventions in methodology or even individual contact with students and teachers who are unmotivated or have low interaction with the school and/or academic community [33]. Furthermore, the evaluation and monitoring practices of students learning is considered as an essential aspect in higher education. Performance monitoring includes assessments and evaluation processes, which play a vital role in providing valuable information that help students, instructor, administrators, and policy makers in higher education institutions to make decisions. The changing factors in contemporary education has led to the use of various data mining techniques to monitor student performance which offers various investigation methods to analyze and discover hidden information in educational systems. The data extracted by such systems continuously comprise the score for certain learning goals so to generate grades and profile learner [2].

2.1 Mining Techniques

According to Mobasher et al. [24], there is no consensus in the literature regarding the ideal data mining technique for each application, not even the criteria to be used for the evaluation of the different data mining techniques. Among the techniques considered efficient in the literature for data mining, are the rules of association that seek to find links between attributes, i.e., based on the assumption that the presence of an attribute in an event implies the presence of another attribute in the same event [29]. The employment of mining with rules of association is appropriate to analyze educational data where it is intended to identify patterns of the learning process and to improve the academic performance of students or teachers in different courses and or school levels, whether in primary, secondary or higher education [24].

In the paper presented by Abu-Oda and Alaa [1], different data mining approaches are applied in order to examine and predict student drop-outs through their university programs. In this study is analyzed 1290 records of students of the course of computer science, between 2005 and 2011. The data collected included a school history, the syllabus of the subjects taught in the first two years of the course, the average student's school, and whether or not the student graduated from the chosen course upon admission to the University. To classify and predict the data of students who dropped out of the course, different classifiers were used in the data set, including Decision Tree (DT) and Bayesian networks. In the paper presented by Baker [7], we identify the application of data association rules to improve the quality of managerial decisions to provide

a quality education, aiming to analyze the data and discover factors that affect the results to increase the chances of success for students and teachers.

2.2 Association Algorithms

Among the association algorithms, the algorithm Apriori, is the algorithm most used to discover rules of association, according to Literature [5]. The Apriori algorithm performs several analyzes on the transaction database, and is able to work with a large number of attributes, resulting in several combinatorial alternatives between them, from successive searches throughout the database and achieving optimum performance in terms of processing [5]. When it is desired to show the frequency and identification of patterns in a data set, the Apriori algorithm is widely used since it presents the frequency of a certain set of selected data. During its application the attributes are tested several times in search of expanded rules which represent patterns of its population [16]. This algorithm works in two steps: in the first step the main frequent item is determined; in the second step the general rule that presents the characteristics of the data set is extracted. The algorithm uses a "bottom up" approach, where frequent subsets are extended once, which is called candidate generation. Each candidate set is tested multiple times until no expanded rule is found [16].

In the use of the algorithm Apriori, the measures that influence the discovery of the rules are: 1. Support that is the percentage of cases where it contains both A and B; 2. Trust that is percentage of cases containing A that also contains B; and 3. Trust is the confidence rate with the percentage of cases containing B [19].

2.3 Related Works

Alom and Courtney [3] made an assessment of the role of student gender on successive rates of educational completion in Australia from primary school Year 1 through to successful completion of high school, and, thereafter, enrolment in university. The results suggested that gender played an important role, especially in some states, and that, in general, enrolment numbers in university in 2016 appeared to be on par with those completing Year 12 in 2015.

The work presented by Fernandes et al. [13] presents an analysis of the academic performance of public school students at the federal district. The authors use federal district education system databases to understand student profiles and their learning styles, in an effort to develop educational policies that improve academic performance and reduce failure rates at the end of each school year. Fonseca and Namen [14] identify factors that relate the profiles and their influences - positive and negative - in students' learning of the mathematics discipline. The authors apply Knowledge Discovery in Databases (KDD) in order to identify the factors that relate the profile of the teachers who teach the Mathematics discipline to the proficiency obtained by their students.

Ashish et al. [12] perform a systematic review of literature from 1983 to 2016 on the clustering algorithm and its applicability and usability in the context of educational data mining. Future insights are outlined based on the revised

literature, and the possibilities for further research have been identified. The work presents that the main advantage of the clustering algorithm application to the data analysis is that it provides a relatively unambiguous learning style scheme to the students, considering a certain number of variables, such as the time spent in completing the learning tasks, group learning, student behavior in the classroom and student motivation for learning.

Asif et al. [4] use data mining methods to study the performance of undergraduate students, focusing on two aspects of their performance. First, predict the academic level of students at the end of a four-year study program. Second, they study typical progressions and combine them with predicted outcomes. Two major groups of students were identified: low performing students and high performing students. The results indicate that by focusing on a small number of courses that are particularly good or bad performance indicators, it is possible to provide timely alerts and support for underachieving students, as well as advice and opportunities for high performance students.

Amirah et al. [28] perform a systematic review of the literature in order to identify the most appropriate methods to predict the performance of students using data mining techniques in educational institutions in Malaysia. Furthermore, they investigate the factors that affect student outcomes in specific courses within the context of Malaysia. The authors propose a model to improve students' performance. The main contribution of the work is to demonstrate how the prediction algorithm can be used to identify the most important attributes in the data of the students to improve their use and success in a more efficient way, using educational data mining techniques, allowing benefits and impacts for students, educators and academic institutions. In the paper presented by [32], analytical approaches to learning, educational data mining and human computer interaction are synthesized to explore the development of more usable forecast models and representations of forecast models using data from a solution environment of collaborative geometry problems: virtual math teams.

3 Method

According to data mining techniques [16,21,23] and educational data mining [7,24], the Apriori algorithm was chosen to support the task of finding association rules in the context of this work and the reference model CRISP-DM [8] to perform the data mining steps, perform tests to evaluate the results found, present results and analyzes, as well as make suggestions based on the conclusions obtained, of possible solutions to implement improvements in the object analyzed with the objective of increasing the satisfaction of teachers, students and the community of the educational system of the analyzed region.

The CRISP-DM reference model defines a set of sequential steps to guide data mining and allows the mining process to be fast, reliable, and with greater managerial control [11]. CRISP-DM involves six phases: domain understanding, data comprehension, data preparation, modeling, evaluation and implementation [19]. The adoption of these phases helps in defining the flows to be used to execute

the mining project. Each phase is structured into several activities, which run cyclically to meet the objectives of data mining [19]. CRISP-DM is a complete and documented methodology. All phases are organized, structured and defined, so the project is easily understood and reviewed [8]. For this study, all CRISP-DM phases were used, except for the implantation phase, which will be performed in our future work.

4 Analysis of Results

Due to the breadth of the analyzed database (11,449.222 lines) and the difficulty in manipulating the data due to its size, the student data file was divided according to the gender of the participants (male - 44.3% of data and female - 55.6% of data). From the analysis of the 99 attributes in the microdata base, it was possible to identify the relevant attributes to answer the Research Questions delimited in this work. The selected dataset is displayed in Table 1. It was used the R-Studio data mining tool [31], which has packages and tools to perform most of the mining steps.

The application of the Apriori algorithm in the national database, containing the records from all Brazil, used the default parameters of the algorithm, e.g., Support in 10% and Trust in 90%. With this application, a set of 272 rules was generated for the data set of the female students and 262 rules for the data set of the male students. Figure 1 presents the rules obtained for the profile of female students in relation to the chosen teaching modality. According to the results, it is possible to conclude that the female students who choose the type of face-to-face teaching were not admitted in higher education through the ENEM and also do not have student funding. Furthermore, the predominant academic degree of this group of students is baccalaureate, the study shift is nocturnal and the administrative category of the place of study is in For-Profit Private Institution. Regarding students who choose distance education, one can only verify that the highest incidence of cases is related to students who were not admitted in higher education through ENEM and that the administrative category of the place of study is Institution of Private for-profit education.

In relation to the profile of male students according to the chosen teaching modality, it was observed that the set of generated rules is very similar to the rules generated for female students, using the same variables, being the similar result for the two groups evaluated. The profile of female students considering the existence of student financing to pay for their studies in higher education, presents the main rules obtained: according to the results presented in Fig. 2 it is possible to identify students who do not have a scholarship to assist in the financing of their studies, declare themselves predominantly as being of white color or race, or have not made the declaration for this variable. Furthermore, the teaching modality is predominantly presential model. Students who do not have student funding have a high incidence of study at night, although some study in the morning. In relation to students who have a scholarship to assist

Table 1. Description of attributes analyzed

Variable name	Description	Description of categories
CATEGORY	Administrative category code	1. Federal Public 2. Public State 3. Municipal Public 4. Private for profit 5. Private non-profit 7. Special
SHIFT	Course shift code to which the student is linked	1. Morning 2. Evening 3. Nightly 4. Integral 5. Distance Education (EaD) (.) Not applicable
GRADE	Code of the academic degree conferred to the graduate by the course	1. Bachelor degree 2. Graduation 3. Technological (.) Not applicable
MODALITY	Course modality code	1. Presential 2. Distance Education (EaD)
COLOR_RACE	Skin Color/Race code	1. White 2. Black 3. Brown 4. Yellow 5. Indigenous 6. Not declared 0. Not declared
ENEM	Informs if the student has enrolled in the course by ENEM	0. No 1. Yes
FINANCING	Informs if the student uses student financing	0. No 1. Yes

in the financing of their studies, most preferred not to declare their color/race, besides the teaching modality being the face-to-face model. Students with student funding also have a high incidence of study at night.

Regarding the profile of male students, considering the existence of funding for higher education studies, as shown in Fig. 3, the following results were obtained: it is possible to identify students who do not have a scholarship to help finance their studies, did not enter higher education through ENEM. These students opted for private non-profit and for-profit institutions, with study hours at night and by the academic degree of baccalaureate and modality of teaching both face-to-face and distance learning. Regarding students who have scholarships to help fund their studies, students declare themselves predominantly as

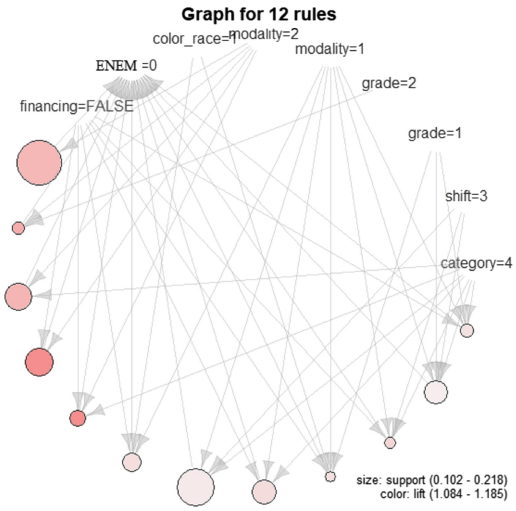

Fig. 1. Evaluation of the profile of female students in relation to the teaching modality.

being of the brown color/race, in addition to the main teaching modality being the face-to-face model and the academic degree being baccalaureate. For those students with student financing, a high incidence of study was observed at night.

Figure 4 presents the main rules obtained for the profile of female students in relation to entering higher education through ENEM. The results show that students who entered through the ENEM or through another form of access have an almost identical profile. These students chose mainly the night shift for study, with little incidence in the rules for the morning study period, the teaching modality is baccalaureate and these students mostly chose not to declare their color/race.

With regard to the male students on the form of entrance into higher education through the ENEM, the Fig. 5 presents the main rules obtained. With the analysis of the result, it is noticed that the identified students declared themselves predominantly as being of the color/race white or brown, are attending the academic degree of baccalaureate, study predominantly in private non-profit and for-profit educational institutions and do not have student financing. Regarding to students who have entered higher education without the ENEM, such as the college entrance examination, their main characteristics are of color/white race and they are studying the academic bachelor degree in the night shift.

With the application of mining classifier algorithms, from the use of the R-Studio tool [31], it was possible to obtain an overview regarding the students' choices according to their gender. Regarding the administrative category chosen by the students, as can be seen in Fig. 6, it is possible to identify that 45% chose private for-profit institutions, 32% study in non-profit private institutions. Approximately 15% chose to study in federal public institutions and 7% in state public institutions. The high choice of private institutions is mainly due to the

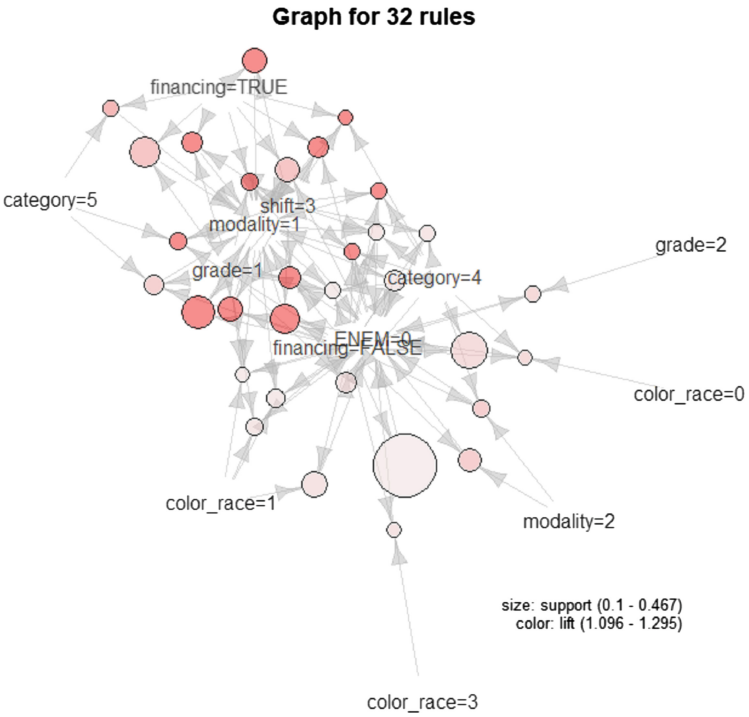

Fig. 2. Evaluation of the profile of female students in relation to the existence of student financing.

greater number of places offered by these institutions. In the classification by gender, a slight predominance by male (15.6%) in federal public institutions is observed when compared to the number of female (12.4%). This predominance is reversed in private for-profit institutions with the distribution of vacancies for male (43.5%) and for female (47.0%).

Regarding the distribution of students by color/race, as shown in Fig. 7, an interesting fact that can be verified is that almost 30% of the students chose not to declare their color/race. Among students who declare their breeding it is noticed that most are of the white color/race (about 37.4%), less than 1% declared themselves indigenous and only about 6% declared themselves black. The variation in all color/race classes between male and female is less than 1%.

Regarding the method of enrollment in higher education, as shown in Fig. 8, the most prevalent was the traditional methods (tests and college entrance examination), being still the most common for students, being around 82%. Only 18% of the students chose to enter higher education through ENEM.

Figure 9 presents the distribution of students according to the existence of student financing. 51.1% of male students and 51.3% of female students declare they do not have student funding, 27.7% of male students and 30.6% of female students declare they do not have student funding and 21.2% of male and 18.2% of female chose not to declare this characteristic.

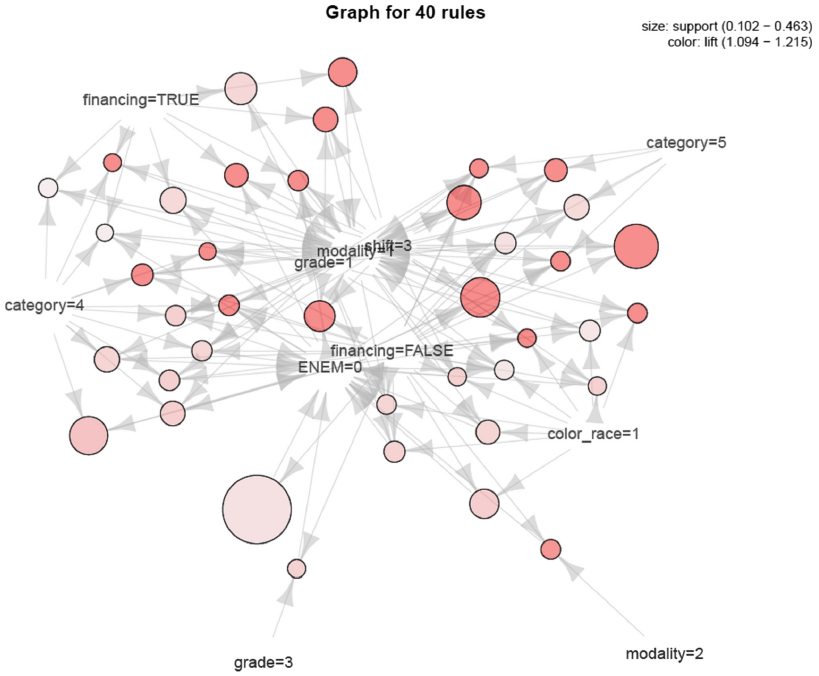

Fig. 3. Evaluation of the profile of male students in relation to the existence of student financing.

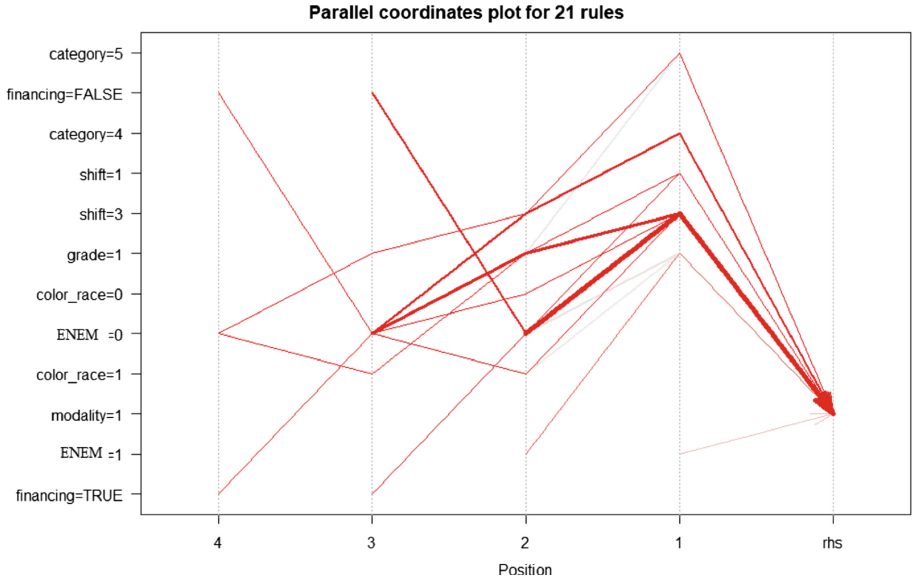

Fig. 4. Profile evaluation of female students who have joined ENEM.

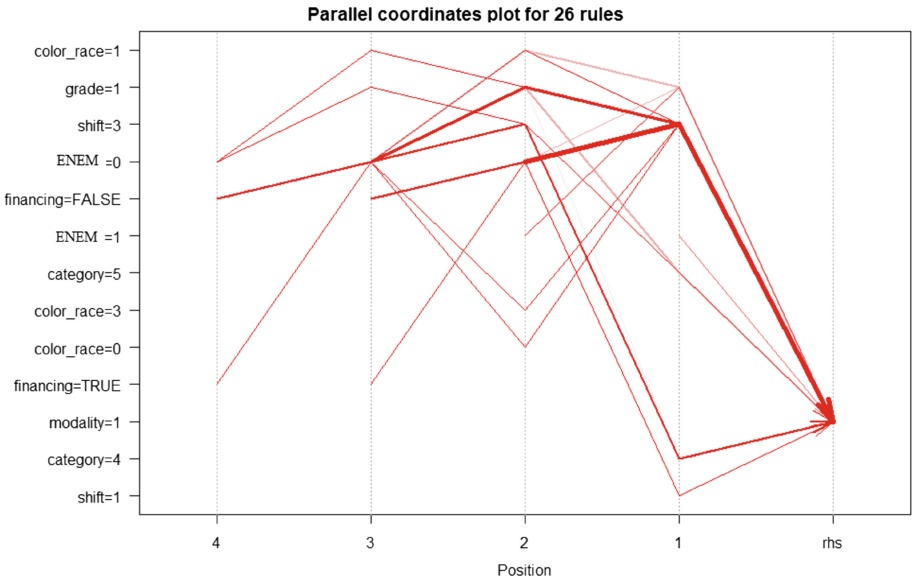

Fig. 5. Profile evaluation of male students who have joined the ENEM.

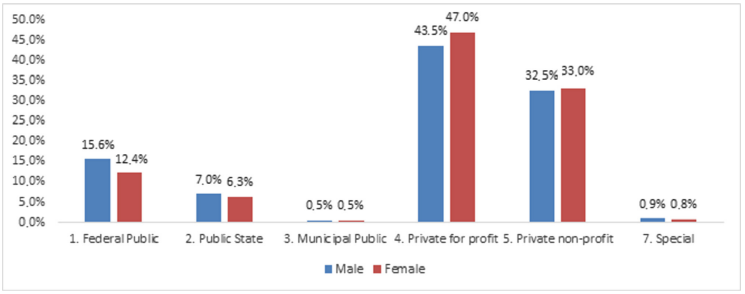

Fig. 6. Distribution of students by Administrative Category.

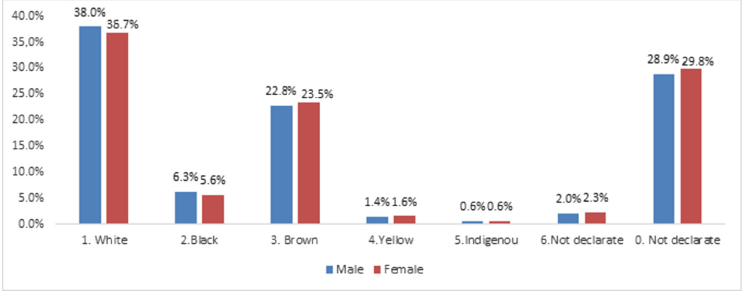

Fig. 7. Distribution of students by color/race.

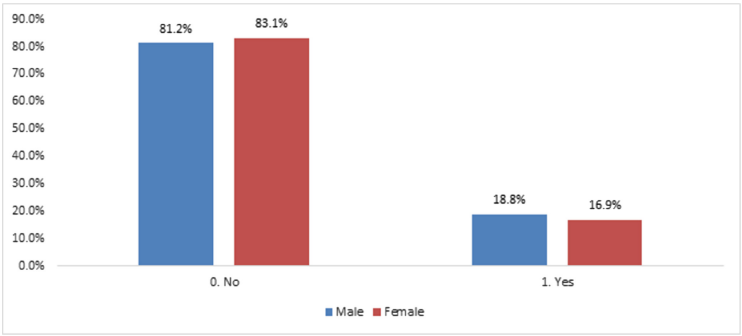

Fig. 8. Distribution of students by the admission method in higher education.

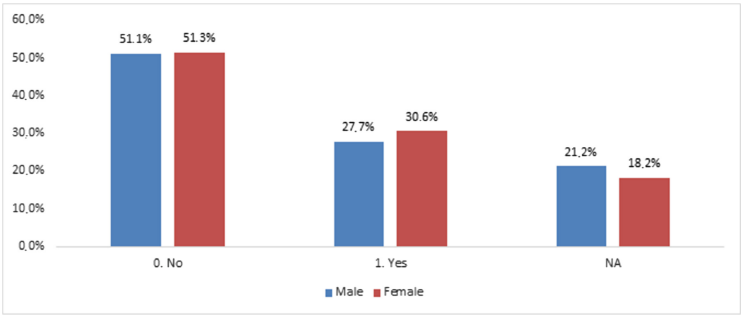

Fig. 9. Distribution of students through the existence of student financing.

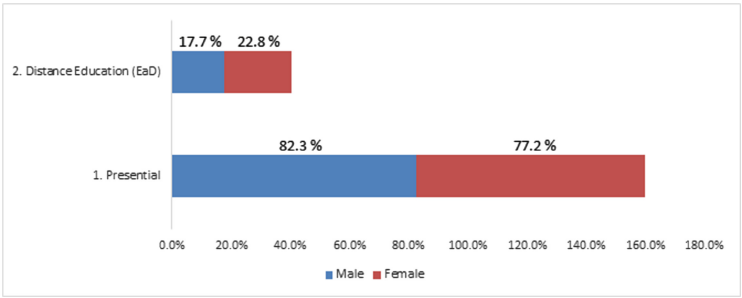

Fig. 10. Distribution of students by Teaching Modality.

The last characteristic analyzed was in relation to the teaching modality. Figure 10 shows the results obtained. It is possible to observe that the face-to-face model is predominant for both male students (82.3%) and female students (77.2%). This is one of the characteristics with the greatest variation in the choice between male and female, surpassing 5% of the total. 17.7% of male and 22.8% of female opted to take a distance learning course.

5 Conclusion

The Higher Education Census organizes and makes available a wide data base related to the data of students, teachers and educational institutions for studies and diagnoses. Analyzing this database, it can be concluded that, although they are an important source of information, they are often not fully exploited due to the inherent difficulties of the large amount of data involved, which is beyond the human capacity to interpret, being necessary the use of suitable tools for its manipulation and analysis.

Given this context, the challenge presented in this study was to make effective use of these data, analyzing them through methodologies that would allow the extraction of information that would provide information for the definition of actions aimed at improving the results of the teaching-learning process of Brazilian Higher Education. The alternative adopted was Data Mining, a technology that is based on statistical concepts and computational intelligence to analyze large volume databases.

It should be emphasized that the scope of this study was restricted to the database of the Higher Education Census conducted in 2016 with the focus on identifying and analyzing the profile of higher education students by assessing their individual characteristics by gender, color/race, teaching modality and student financing.

The results obtained from the data mining made possible the profile description of the Brazilian higher education student in relation to the different types of teaching offered (face-to-face or distance education), allowing to identify that more than 80% of Brazilian students chose modality of face-to-face teaching. Furthermore, it allowed investigating the distribution of scholarships for student funding by different groups of students and evaluating the distribution of places occupied in higher education according to color/race and gender informed by the students.

As a continuation of this work, it is expected to evaluate data on higher education institutions, available courses, the profile of teachers and characteristics of the locations covered by Brazilian higher education institutions. This data set is also made available by the Census and if studied together with the other information can identify patterns of behavior and propose actions to further assist in the improvement of the educational system.

Acknowledgment. This research work has the support of the Research Support Foundation of the Federal District (FAPDF).

References

1. Abu-Oda, G.S., El-Halees, A.M.: Data mining in higher education: university student dropout case study. Int. J. Data Min. Knowl. Manage. Process **5**(1), 15 (2015)
2. Aldowah, H., Al-Samarraie, H., Fauzy, W.M.: Educational data mining and learning analytics for 21st century higher education: a review and synthesis. Telematics Inform. **37**, 13–49 (2019)
3. Alom, B., Courtney, M.: Educational data mining: a case study perspectives from primary to university education in australia. Int. J. Inf. Technol. Comput. Sci. (IJITCS) **10**(2), 1–9 (2018)
4. Asif, R., Merceron, A., Ali, S.A., Haider, N.G.: Analyzing undergraduate students' performance using educational data mining. Comput. Educ. **113**, 177–194 (2017). https://doi.org/10.1016/j.compedu.2017.05.007
5. Awan, S., Dadan, R.: Graduate rate analysis of student using data mining and algorithm apriori. Int. J. Soft Comput. **12**(5), 287–293 (2017)
6. Baker, R., et al.: Data mining for education. Int. Encycl. Educ. **7**(3), 112–118 (2010)
7. Baker, R.S., Inventado, P.S.: Educational data mining and learning analytics. In: Larusson, J.A., White, B. (eds.) Learn. Anal., pp. 61–75. Springer, New York (2014). https://doi.org/10.1007/978-1-4614-3305-7_4
8. Barclay, C., Dennis, A., Shepherd, J.: Application of the crisp-dm model in predicting high school students' examination (CSEC/CXC) performance. In: Knowledge Discovery Process and Methods to Enhance Organizational Performance, p. 279 (2015)
9. Canedo, E.D., Santos, G.A., de Freitas, S.A.A.: Analysis of the teaching-learning methodology adopted in the introduction to computer science classes. In: 2017 IEEE Frontiers in Education Conference (FIE), pp. 1–8. IEEE (2017)
10. Chakraborty, B., Chakma, K., Mukherjee, A.: A density-based clustering algorithm and experiments on student dataset with noises using rough set theory. In: 2016 IEEE International Conference on Engineering and Technology (ICETECH), pp. 431–436. IEEE (2016)
11. Chapman, P., et al.: CRISP-DM 1.0 step-by-step data mining guide. CRISP DM consortium (updated 2010) (1999) (2000)
12. Dutt, A., Ismail, M.A., Herawan, T.: A systematic review on educational data mining. IEEE Access **5**, 15991–16005 (2017)
13. Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., Van Erven, G.: Educational data mining: predictive analysis of academic performance of public school students in the capital of brazil. J. Bus. Res. (2018). Elsevier
14. Fonseca, S.O.D., Namen, A.A.: Data mining on INEP databases: an initial analysis aiming to improve Brazilian educational system. Educação em Revista **32**(1), 133–157 (2016)
15. Hung, J.L., Crooks, S.M.: Examining online learning patterns with data mining techniques in peer-moderated and teacher-moderated courses. J. Educ. Comput. Res. **40**(2), 183–210 (2009)
16. Jang, S., Park, K., Kim, Y., Cho, H., Yoon, T.: Comparison of H5N1, H5N8, and H3N2 using decision tree and apriori algorithm. J. Biosci. Med. **3**(06), 49 (2015)
17. Jha, J., Ragha, L.: Educational data mining using improved apriori algorithm. Int. J. Inf. Comput. Technol. **3**(5), 411–418 (2013)
18. Kabra, R., Bichkar, R.: Performance prediction of engineering students using decision trees. Int. J. Comput. Appl. **36**(11), 8–12 (2011)

19. Kalgotra, P., Sharda, R.: Progression analysis of signals: extending CRISP-DM to stream analytics. In: 2016 IEEE International Conference on Big Data (Big Data), pp. 2880–2885. IEEE (2016)
20. Kavitha, G., Raj, L.: Educational data mining and learning analytics - educational assistance for teaching and learning. CoRR abs/1706.03327 (2017)
21. woo Kim, C., Ahn, S.H., Yoon, T.: Comparison of flavivirus using datamining-apriori, k-means, and decision tree algorithm. In: 2017 19th International Conference on Advanced Communication Technology (ICACT), pp. 454–457. IEEE (2017)
22. Lara, J.A., Lizcano, D., Martínez, M.A., Pazos, J., Riera, T.: 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. Comput. Educ. **72**, 23–36 (2014)
23. Luna, J.M., Padillo, F., Pechenizkiy, M., Ventura, S.: Apriori versions based on mapreduce for mining frequent patterns on big data. IEEE Trans. Cybern. **99**, 1–15 (2017)
24. Mobasher, G., Shawish, A., Ibrahim, O.: Educational data mining rule based recommender systems. In: CSEDU (1), pp. 292–299 (2017)
25. Paiva, R., Bittencourt, I.I., Lemos, W., Vinicius, A., Dermeval, D.: Visualizing learning analytics and educational data mining outputs. In: Penstein Rosé, C., Martínez-Maldonado, R., Hoppe, H.U., Luckin, R., Mavrikis, M., Porayska-Pomsta, K., McLaren, B., du Boulay, B. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 251–256. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93846-2_46
26. Papamitsiou, Z., Economides, A.A.: Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence. J. Educ. Technol. Soc. **17**(4), 49–64 (2014)
27. Ray, S., Saeed, M.: Applications of educational data mining and learning analytics tools in handling big data in higher education. In: Alani, M.M., Tawfik, H., Saeed, M., Anya, O. (eds.) Applications of Big Data Analytics, pp. 135–160. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76472-6_7
28. Shahiri, A.M., Husain, W., et al.: A review on predicting student's performance using data mining techniques. Procedia Comput. Sci. **72**, 414–422 (2015)
29. da Silva, L.A., Peres, S.M., Boscarioli, C.: Introdução à mineração de dados: com aplicações em R. Elsevier Brasil (2017)
30. Silva, R., Ramos, J.L.C., Rodrigues, R., Gomes, A.S., Fonseca, A.: Mineração de dados educacionais na análise das interações dos alunos em um ambiente virtual de aprendizagem. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), vol. 26, p. 1197 (2015)
31. Studio, R.: RStudio: integrated development environment for R, p. 74. RStudio Inc., Boston (2012)
32. Xing, W., Guo, R., Petakovic, E., Goggins, S.: Participation-based student final performance prediction model through interpretable genetic programming: integrating learning analytics, educational data mining and theory. Comput. Hum. Behav. **47**, 168–181 (2015)
33. Yukselturk, E., Ozekes, S., Türel, Y.K.: Predicting dropout student: an application of data mining methods in an online education program. Eur J. Open Distance E-learn. **17**(1), 118–133 (2014)