Undergraduate Students' Effectiveness in an Institution With High Dropout Index

1st João Lucas Oliveira
Instituto de Informática
Universidade Federal de Goiás
Goiás, Brazil
joaooliveira@discente.ufg.br

2nd Ana Paula Ambrósio
Instituto de Informática
Universidade Federal de Goiás
Goiás, Brazil
apaula@inf.ufg.br

3rd Uyara Silva
Instituto de Informática
Universidade Federal de Goiás
Goiás, Brazil
uyara.silva@ifg.edu.br

4th Jacques Brancher
Instituto de Informática
Universidade Estadual de Londrina
Paraná, Brazil
jacques@uel.br

5th Jacinto José Franco
Instituto de Informática
Universidade Federal de Goiás
Goiás, Brazil
jacinto.franco@bag.ifmt.edu.br

***Abstract*—Currently, the dissemination of open data in conjunction with Educational Data Mining (EDM), learning analytics, e-learning, intelligent systems, intelligent tutors, and online judge techniques have had made useful contributions to the field of education through the knowledge generated from data analysis. Identifying factors that allow us to understand how students learn and their behavior has aided managers and teaching professionals to identify the best teaching settings. This study aims to do a comparison of academic success with other studies in the literature. Two thousand four hundred ninety-nine students were analyzed for over 11 years. These students belong to a Brazilian university and three undergraduate courses of computing (Computer Science, Software Engineering and Information Systems). The Statistical and data mining techniques were used to extract information that can validate the hypotheses of this study. Our Main objective is to seek which factors tend to contribute to students' retention, dropout, difficulties, and academic success. For reach this objective, we compare gender effectiveness and course curriculum grade. The data showed that some factors, not previously analyzed by other studies, tend to influence student performance.**

***Index Terms*—Academic success, Computer education, EDM**

I. Introduction

The Educational Data Mining (EDM) describes a research field concerned with the application of data mining, machine learning and statistics to information generated from educational settings. As saw in [1]the field seeks to develop and improve methods for exploring this data, in order to discover new insights about how people learn in the context of such settings. The field of EDM is closely linked to that of learning analysis, educational psychology, and learning science.However, it is essential to emphasize that the development of EDM is mainly due to data initiatives available for analysis, which currently has generated debates and discussions about the availability of such data.

In 2016, Brazil adopted a measure to make the governmental data available to the community, along with the university data as well. This measure aims to promote dialogue between actors in society and with the government to think about the best use of data, promoting positive impacts on the social and economic sphere. [2] The main stakeholders in these data are the surveyed and educational managers. In addition, to promoting the transparency of public data, these data can be compared to other teaching models globally [3].

The incentive to publish data is the development of public policies, security guidelines, and definitions regarding the intellectual property of the data to be published and reused. The Brazilian reality points to the need to broaden the discussions about the availability and use of these data. The Brazil currently has three fundamental attitudes for the dissemination of this information, e.g: The LAI - Law on Access to Information, Law No. 12.527/2011, regulates the constitutional right of access to public information; The decree No. 8.777/2016, which requires the administration Brazilian public to make data available in a format open to improving the culture of transparency; The Law General Personal Data Protection Act (LGPD) No. 13,709 regulates the processing of personal data of customers and users by public and private companies that will be valid in 2020.

There's no question that the educational sphere is one of the great beneficiaries and stakeholders of public data democratization. About 6% of gross domestic product (GDP) is invested in public education in Brazil, higher than the Organisation for Economic Co-operation and Development (OECD) average (5.5%), which includes major world and peer economies such as Argentina (5.3%), Colombia (4.7%), Chile (4.8%), Mexico (5.3%), and the United States (5.4%). Brazil surpasses 80% of the countries, including several developed countries in spending on GDP education[4]. Nevertheless, in PISA 2018, students in Brazil scored lower than the OECD average in reading, mathematics, and science. Only 2% of students performed at the highest levels of proficiency (Level 5 or 6) in at least one subject (OECD average: 16%), and 43% of students scored below the minimum level of proficiency

(Level 2) in all three subjects (OECD average: 13%) [5].

The MEC recently released a technical note in February 2018, which analyzes data from 63 federal educational institutions in the period 2009 and 2016 [6]. According to the report, a federal university student in 2016 cost R$ 3,129 per month or R$ 37,548 per year. The calculations include data from Capes, Censo do Ensino Superior, and the Tesouro Nacional with inflation correction. Believes that the MDE can help find the main causes that evidence retention and evasion.

Some authors, like Baker [7], demonstrate the opportunities that Brazil has to improve education through Educational Data Mining (EDM). Several initiatives have been set up in universities to use mining techniques to extract educational indicators from their data [8]. Furthermore, as in [9], [10], [11], the situation of dropping out, student retention, and academic success are recurring themes in research involving academic success in EDM [12]. Successful academic research is an excellent opportunity to measure the quality of teaching and generate savings on education spending, especially at the undergraduate level, where more mature research and technological development are found. The Data Mining area has significantly contributed to the field of academic success. One of the factors of this contribution is Artificial Intelligence (AI) that has been acting in different areas of knowledge, optimizing, and analyzing processes.

This study analyzes academic data from undergraduate Science Computer core students at one of the top 20 universities in Brazil, looking for factors that decrease the chances of academic success. We intend to promote the analysis of these data and compare the results found with other works within the literature. Dissemination of this information can help researchers and teaching specialists to develop studies centered on the factors that affect student education.

II. Related Work

The area of educational data mining has been an essential initiative for pedagogical and administrative management. The constant search for new techniques and methods to understand students' academic environment and learning configurations has been a keynote address by one of the most public entities in this area The International Educational Data Mining Society [13]. In [14] a recent literature review of EDMs analyzes the clustering techniques were used to group similar works in an attempt to group the types of research being developed in EDM. Among the highest indices are: General Educational Data Mining, E-Learning, and Learning Style. This study was analyzed works from the last 30 years. Research such as [1], [15], and [16], are excellent sources of how the field of EDM has developed. However, as can be seen in these works, the focus of EDM analysis has always been centered on the student, their behaviors and patterns.

A. EDM on university data

The application of EDM in university data provides us with indicators of the quality of public education, besides projecting the necessary demands that the labor market needs. In [17], the authors build a progressive analysis of students from elementary school through high school to university enrollment. The authors used mining and statistical techniques to generate insights into how students' academic progression develops.

The study of [18] has investigated three research questions. The first question concerns predicting students' performance using marks only, no socio-economic data(was possible with reasonable accuracy). The second question strives for deriving courses that can serve as effective indicators of good or poor performance in the degree program(about 4 courses was identified). The third question involves investigating how students' academic performance progresses over the four-year degree (was identified that low-performance students tend to still a low-performance across the course). Modelos de machine learning e statistics was used to answer the questions.

In [19] the authors share the experiences gained in analyzing student dropout risk through school data, sociodemographic data and learning management systems (LMSs). The algorithm used is based on Random Forest. The authors discuss how important the data experiments were to size the main features that covered the search space for their research questions. Data generated by the LMS increased the accuracy percentage of the models, when used.

Already in [20] the authors try to predict student academic success through mathematical models. However, the great contribution of the authors is the demonstration of the variables that determine student academic success, such as Calculus and physics. Another very important study is also the [21] who use graph mining techniques to find retention patterns in undergraduate curricula of Computer Science. In this study it is again evidenced that mathematical theoretical subjects tend to have high failure rates, generating student retention. This technique is an area of EDM that has been growing and showing behavior patterns in physical environments (classrooms) and virtual environments (e-learning, Forum, distance learning courses), more about this area can be seen in studies of [22].

B. Academic success

In [23] factors that promote student academic success are identified, various exploratory and inferential statistical analyzes have been applied. As a result, the high school leaving grade is by far the best predictor of both the probability of graduating and the final grade obtained at university. Other factors, gender or social origin, play only a minor role. Moreover, high school performance is strongly linked to whether a student graduates at all from university than to the success in a specific faculty. The results differences emerge among the various faculties. The same is investigated in [24] but with focus on three factors: average grade (GPA), number of credits (ECTS) and intention to persist with 243 first year college students in the Netherlands.

In [25] the authors use complex graph and network techniques to identify behavioral aspects concerning student interactions. They indicate that integration in social networks is a critical success factor in academic settings and contradict arguments that friendships at school and college are potentially disruptive for academic development [26].

On the other hand, the authors [27] investigated the university in London, the United Kingdom to know which factors tend to promote the success academic by the words of participants. The participants defined academic success e.g: the accomplishment of the learning process; gaining subject knowledge; developing employability skills.

Stimulating critical thinking, keeping it motivated, and engaged is one of the challenges of academic success. In [28], the author's analysis of constructivism has led him to conclude that the epistemology of computer science is significantly different from that of, say, physics. Nevertheless, the basic tenet of the theory (that knowledge is constructed by the student) applies to computer science, and its central implication is that models must be explicitly taught.In most computer courses students are encouraged to create applications and build abstract solutions computationally for real problems.

III. Methodology

For this study we will use research based on investigations about factors that influence retention and dropout problems in undergraduate courses. It is a comparative exploratory research among works that also investigate such causes in order to investigate whether factors exposed in the literature in different regions tend to manifest themselves in the same way in the present study. In addition, we will advance in what the authors expressed as "future works" in order to explore behavioral characteristics that demonstrate retention and avoidance. That said, the main hypotheses investigated in this study are:

1) Are the frequent causes of inactivation shown to be equivalent to those seen in the literature?
2) Are the disciplines found to be more difficult in this study also present in the literature?
3) Does the choice of teacher for discipline directly influence student performance?
4) Does the difference between admission periods tend to be a determining factor for student performance?
5) Does the gender difference between students in the STEM group characterize a difference in performance?

The University Federal de Goiás today has a prominent problem with student retention. A few students are graduating, many students giving up, and being held back in the university. This study will investigate which factors tend to contribute to these causes. The database is organized by a total of 2499 records by students within three Computer Science programs e.g: Computer Science (977 students), Software Engineering (564 students), Information Systems (958 students). The records were collected by an internal management system of the Federal University X in .CSV format. The data were collected from 2009 to half of 2019 and contain information about the students' academic attributes, i.g: name, course, discipline, score, global score, relative score, period, teacher name, code unique, date of birth, number of dropout, and 37 others attributes. For our analysis we use jupyter-notebook as data processing environment. The main libraries used were Python 3.7, Scipy, Pandas, Matplotlib and Seaborn.

IV. Results

The figure 1 show the overview of Institute of Computer Science and its possible verify that among the 3 analyzed courses the smallest student admits is the Software Engineering program, and the largest is the Computer Science program. However, it should be noted that the Computer Science program is the oldest of the programs and also the Software Engineering program has only one student admission process per year, unlike the other two courses which have 2 admission processes for each year.

The situation of students in the programs reflects what is known in the literature. The rate of students who will graduate in the period does not exceed 30%. On other hand in [29] studies report that the percentage of graduates for each class is approximately 40%. The courses analyzed are part of what is known in the literature as STEM (Science, Technology, Engineering and Mathematics) which have the highest dropout rates and the lowest pass rates [30]. In addition, most students are active and excluded, which indicates that most students are passing the time to graduate, which generates extra expenses in University budgets. The student retention and dropout problems have been widely studied in the education field, as it's one of the most sensitive and complicated problems to solve [31], [32].

Based on the monetary metrics discussed in the introduction. If we calculate the amount invested in students of the 3 programs, we have a total of 904 dropout students, which may have generated a loss statements of R$ 33.943,392 for the university. The number increases even more if we consider all active students who did not complete on time, all programs within a university, and the 63 federal institutes that were analyzed in the MEC survey.

The figure 1 also shows the total average score students, average score graduate degree students, and average score not graduate degree students, some analyses on our database have shown that a student with poor performance tends to continue with poor performance, and students with good performance tend to continue to perform well. The university offers some means to help students who have difficulty with certain subjects, such as mentoring, summer and winter courses, and class monitors, but the results are not encouraging.

There is a slight difference between the average scores of male and female students, except for the Information Systems program. For better analysis statistical tests (T-student test) were made to verify if there is significant difference between genders. Firstly, the data were normalized and ordered for test application. It was identified that in the CC and SE programs there is a significant difference between the means score of students with respective p-value, 0.06 and 0.03, but in the IS course there was no significant difference, p-value 0.12. The experimenter used the global average score of male and female students, the amount of grades were balanced for the same amount of both male and female average scores. The

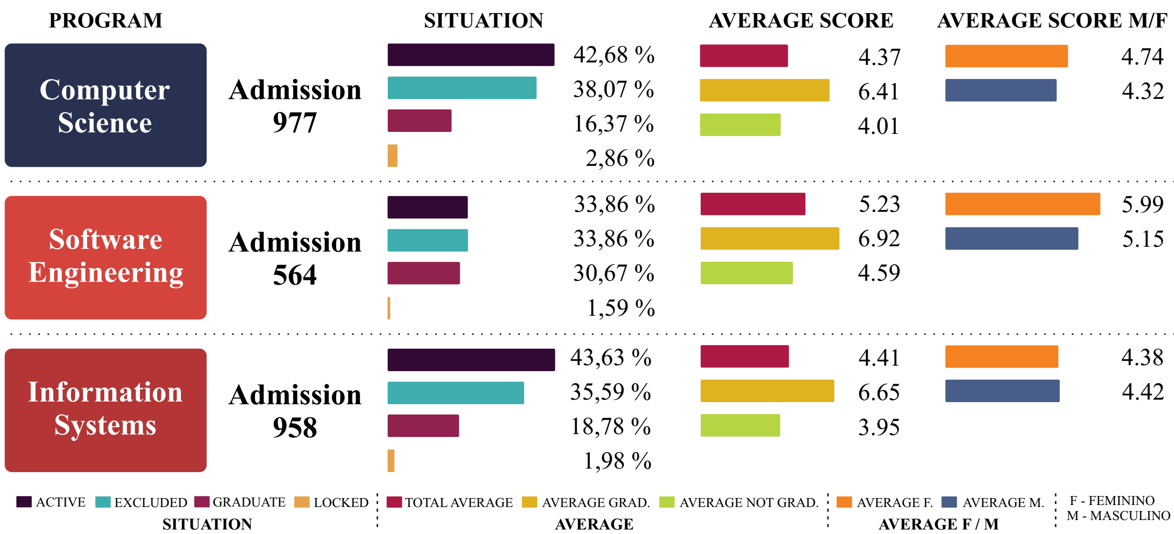

Figure 1. Overview of undergraduate students between 2008 and 2019.1 with status and average of the 3 programs offered e.g: Computer Science, Software Engineering and Information Systems. The term *AVERAGE NOT GRAD* it means undergraduate students who not complete the course. the same goes for *AVERAGE GRAD* to concluded students.

total 1000 test interactions were made to validate the value of p-value.

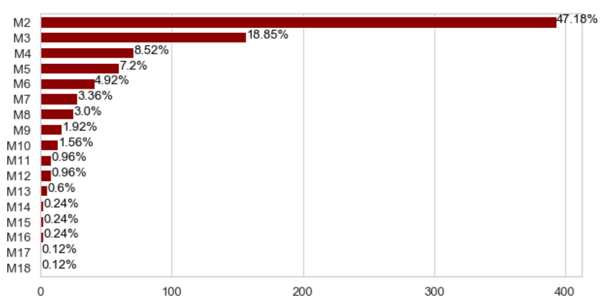

Figure 2. Most frequent inactivation of the three programs between 2008 and 2019.

Figure 2 M2 The most representative of the causes of inactivations is referring to "Excluded for not renewing ties with the university," M3 refers to "Dropout," in M4 "Exclusion for failure and low score" M5 "chosing another program" is the most common cause of dropout. In most cases, students either drop out of the course or are excluded by excessive numbers of failures. The other reasons refer to specific disapproval of the subject, discipline, and deadlines regarding courses.

The table I presents a survey regarding the students' training time. It may seem a bit odd, however the cases where Min = 1 in computer science refers to students who took advantage of the grade and graduated in a short time. Training time on average students delay 1.5 to graduate, which may be one of the institute's retention indicators.

Table I

AVERAGE TIME OF GRADUATION OF COMPUTER SCIENCE STUDENTS BETWEEN 2008 AND 2019. THE NOTATIONS "STD" REFERS TO THE STANDARD DEVIATION, "MIN" AND "MAX" TO MINIMUM AND MAXIMUM GRADUATION TIME.

Program	Mean(years)	Min.	Max.	Std
Computer Science	$\simeq 5.5$	1	10	1.45
Engineering Software	$\simeq 5$	4	10	1.09
Information Systems	$\simeq 5.5$	2	10	1.41

The table II shows the failure rate of the subjects that were taken by students throughout the undergraduate process. The normalization techniques used in the table below obey the criterion of ratio between the total number of subjects taught by the number of failures and approvals. With a variation between 0 to 100% always following the criteria of equivalent disciplines. The data processing included: i) Identifying and treating records with null attributes, ii) identifying and treating noisy data, ii) records with missing data greater than 60% were excluded. Only the subjects that were taught with more than 100 students were included in this calculation. We chose this limit based on the time of possible classes and their variation of teachers during the courses. It is possible that up to three teachers can teach this subject before being taught 100 students, but in many of the observed cases the difference in failure between teachers was not evident before taken by 100 students, so we defined this threshold. The records of students belonging to the sample constitute: graduating students, students who transferred, locked, escaped or dropped out of the course.

Table II
THE SUBJECTS WITH THE HIGHEST FAILURE RATE DURING 2008 AND 2019.1

COMPUTER SCIENCE	%	ENGINEERING SOFTWARE	%	INFORMATION SYSTEMS	%
IME0286	68.38	IME0075	67.64	INF0285	69.44
IFI0080	63.36	INF0285	57.39	IME0075	67.01
IME0075	62.98	IME0286	39.70	INF0286	57.65
IME0164	59.08	INF0186	36.67	INF0135	48.62
IME0006	53.68	INF0284	36.15	INF0284	41.59
INF0285	51.86	INF0015	36.05	INF0016	39.09
IME0078	50.29	INF0011	33.06	INF0203	38.37
IME0073	48.50	INF0194	31.10	INF0163	37.70
INF0204	48.12	INF0022	30.85	INF0053	36.96
INF0127	46.29	INF0189	30.47	INF0061	36.93

The names of the subjects were coded. Among the top 10 of failure rate in the CC program 80% of them are from the mathematics core (for this we also consider the subjects of physics), on EG program are 30%, and on SI program are 40%, the same results could be observed also in the studies of [21]. The study of (Costa and Bernardini et. al.) [21] graphs were used to find the disciplines that most difficult the students' process of graduation. The results showed that among the most costly paths found, the subjects that involved mathematics promoted the highest retention rate. When we analyze the 2 subjects that have the highest failure rate in both programs we show that 100% belong to the core of mathematics.

Another hypothesis investigated is regarding the period of dropout of students. We investigated the period of the highest dropout rate of students. In the CC program, there are 45.57 % of students drop out in the first year of study, in ES 41.66 %, in SI 47.50 %. When looking at the first two years, the numbers are even more alarming, in CC 70.49 %, ES 57.69 %, and SI 66 %. It is known in the literature that most students tend to drop out in the first year, which is reflected in this analysis.

The data from the table II were considered only the subjects that have a total of 100 students attended. This guaranteed us that the subjects in the study were just those with a comprehensive class time. Among the programs the computer science course contains more mathematical disciplines than the other courses. The mathematical basis in this course is more present, and this may be probably influencing the outcomes.

During the database investigation, it was suspected that some subjects taken by different teachers would have different approval and fail rates. We investigated among the subjects shown in table II that demonstrate the subjects with the highest failure rate. The *Physics I* and *Discrete Mathematics* have been shown to vary widely from teacher to teacher. Likewise, for some teachers with high pass rates than others, in some cases like the example, *NF0008P* from table III has a much more significant variation in the pass rate than teacher *NF00007P*. Therefore, the teaching method used by teachers with high failure rates should be reviewed.

Table III
INDICATIVE OF TEACHER FAILURE RATE

COURSE	INSTRUCTOR	APPROVAL %	FAILURE %
IME0286	NF0001P	2.38	**97.62**
	NF0002P	55.1	44.9
	NF0003P	61.36	65.91
INF0182	NF0004P	28.02	**71.98**
	NF0005P	50.99	49.84
	NF0006P	31.65	67.14
INF0186	NF0007P	55.83	63.6
	NF0008P	**80.93**	19.07

Figure 3 describes the survey on the situation of students subdivided by gender. The female participation does not exceed 15% in both courses. Besides, in [33], the authors show a graph of women applicants to computer science compared to other HE courses - all UK universities the number of women who applied to computer courses is also 15%, but this is the number applications and not for entry. In [34] the authors speak their thesis that girls, even those who grew up with technology in their homes and took advanced math classes in high school, less likely than boys to pursue computer science and engineering. They argue that women's choices are constrained by societal factors, particularly their stereotypes about the kind of people, the work involved, and the values of these fields. These perceptions, even if they are not accurate, shape the academic choices that girls make by communicating to them where they belong. Moreover, To investigate the following hypothesis, "Do the average score female students tend to be greater than the male"? For this, we analyzed the situation of the students. The percentage of female admissions/graduates tends to be higher in CC and IS courses, with a slight difference in SE. Furthermore, The overall average score of female students is higher in the CC and ES courses, with a slight difference in the IS course.

The figure 4 is a graph referring to the failure rate among students who entered the first and second semesters of the computer science course. There is a considerable difference between the period from 2009 to 2012, having a growing in 2013 and again varying from 2015. For this analysis, we investigated the hypotheses that students entering the first half-semester ones tend to perform better than students entering the program in the second half-semester. This is since students entering the first one are mostly students with higher scores in the ENEM (National High School Exam), which is the gateway to most of the federal universities in the country. The metrics found are Computer Science course (differences score average: 4.75, global average: 31.04, differences variance: 9.0, standard deviation: 3.0), but Information Systems course (differences average: 3.21, global average: 22.10, differences variance: 4.38, Standard deviation: 2.09). What we consider as "mean, variance and standard deviation of differences" is related to the difference in values between the first and second semesters. Does not analyze the software engineering course by the fact that in this program, or entry only occurs in the second semester, therefore, the data is not required for

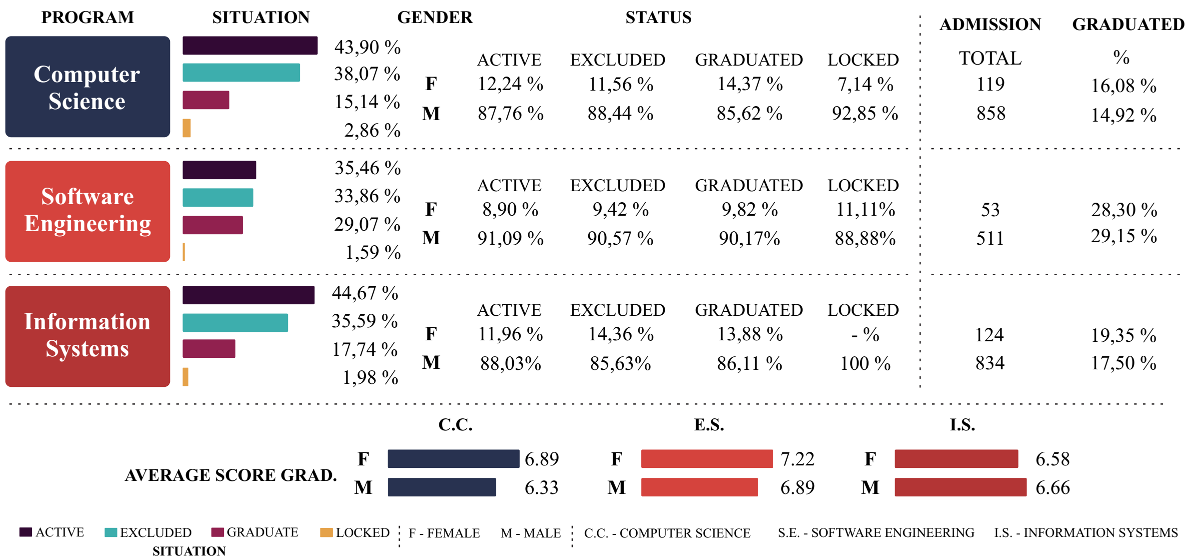

Figure 3. The situation of students subdivided by gender and graduation rate by gender between the period 2008 and 2019.1.

analysis.

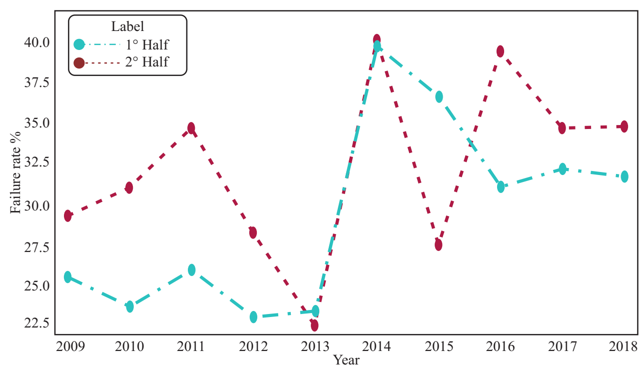

Figure 4. The difference in performance between first and second-semester students between 2009 and 2018 of Computer Science program.

In addition, Computer science courses have the worst ratings of mean, variance, and standard deviation, but it needs to be stricter about this assessment. Investigating our hypothesis that there is a significant difference in failure rates between the first and second semesters, we applied a Student t-test to validate the results. First, we sort and normalize the values and apply the Test-t. In the Computer Science course, we had (statistic = 0.2348, p-value = 0.8169), so there is no significant difference between first and second-semester failure rates. In the Information Systems program, the values found were (statistic = 0.3089, p-value = 0.7609), demonstrating that there is no significant difference in the Information Systems program either.

V. Conclusion

The first step to a proper data analysis in higher education is to ask the right questions. Only awareness of what is being sought can lead to the final answer. However, in the academic world, several factors must be taken into account, from knowledge of epistemology, referring to the theory of knowledge sought; Ontological, regarding the content of knowledge obtained and deontological regarding the principles and practices adopted by the researched. The analysis of the students shows the learning needs of students, facilitating the prescription of teaching appropriate to the demands of the cohort. As seen in the analyzes, there is a difference, especially in the courses with a higher failure rate between teachers. It is necessary to rethink whether universities have evolved technologically and pedagogically; and if their competent entities have been following such evolution. Another factor is the prediction of student performance, which allows us to see if current actions are promoting a real improvement in learning. The overall course averages still very low than expected for undergraduate courses, and quite close to STEM courses in literature. Gender differences, especially in STEM courses, have been the subject of extensive debate in the literature. As seen, there is a difference in input, grade, and formation about gender, especially in some cases, this difference is not significant. However, the number of female admissions still does not exceed 20%, studies such as [35], [36], [37], [38] provide more deeper debate on the factors influencing women's professional and social career.

The analyzes made in this study are essential to measuring the evolution and trends of higher education and the importance of open data as a measure of monitoring and evaluation of data disclosed for comparison between pedagogical organizations. The programs analyzed in this study belong to the core of computing, where they had not had good results for some years. The average number of failures in subjects per semester is 2, which may demonstrate poor primary education, immaturity in the face of a new environment change from necessary to higher education, or even low student motivation.

Acknowledges

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

[1] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *JEDM| Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[2] C. K. P. Pereira, S. W. S. Siqueira, B. P. N. Nunes *et al.*, "Dados conectados na educação," in *Congresso da Sociedade Brasileira de Computação-CSBC*, 2017.

[3] W. Alcantara, J. Bandeira, A. Barbosa, A. Lima, T. Ávila, I. Bittencourt, and S. Isotani, "Desafios no uso de dados abertos conectados na educaçao brasileira," in *Anais do DesafiE-4o Workshop de Desafios da Computação Aplicada à Educação. CSBC*, 2015.

[4] T. Nacional, "Aspectos fiscais da educação no brasil," 2018. [Online]. Available: "http://www.tesouro.fazenda.gov.br/documents/10180/318974/EducacaoCesef2/eb3e416c-be6c-4325-af75-53982b85dbb4"

[5] OECD, "Programme for international student assessment (pisa) 2018, brazil," 2018. [Online]. Available: "https://www.oecd.org/pisa/publications/PISA2018_CN_BRA.pdf"

[6] MEC, "Nota técnica mec se n 4/2018," 2018. [Online]. Available: "http://forplad.andifes.org.br/sites/default/files/forplad/comissaoplanejamento/NT_04-2018_e_anexos__apura%C3%A7%C3%A3o_do_custo_das_universidades.pdf"

[7] R. Baker, S. Isotani, and A. Carvalho, "Mineraçao de dados educacionais: Oportunidades para o brasil," *Brazilian Journal of Computers in Education*, vol. 19, no. 02, p. 03, 2011.

[8] M. Kumar, A. Singh, and D. Handa, "Literature survey on educational dropout prediction," *International Journal of Education and Management Engineering*, vol. 7, no. 2, p. 8, 2017.

[9] M. P. Martins, V. L. Miguéis, D. Fonseca, and A. Alves, "A data mining approach for predicting academic success–a case study," in *International Conference on Information Technology & Systems*. Springer, 2019, pp. 45–56.

[10] V. N. Uzel, S. S. Turgut, and S. A. Özel, "Prediction of students' academic success using data mining methods," in *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2018, pp. 1–5.

[11] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil," *Journal of Business Research*, vol. 94, pp. 335–343, 2019.

[12] R. K. Rambola, M. Inamke, and S. Harne, "Literature review-techniques and algorithms used for various applications of educational data mining (edm)." in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. IEEE, 2018, pp. 1–4.

[13] EDM, "Educational data mining," 2019. [Online]. Available: "http://educationaldatamining.org/"

[14] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15 991–16 005, 2017.

[15] M. P. Martins, V. L. Migueis, and D. Fonseca, "Educational data mining: A literature review," in *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2018, pp. 1–6.

[16] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[17] B. Alom and M. Courtney, "Educational data mining: a case study perspectives from primary to university education in australia," *Int. J. Inf. Technol. Comput. Sci.(IJITCS)*, vol. 10, no. 2, pp. 1–9, 2018.

[18] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education*, vol. 113, pp. 177–194, 2017.

[19] A. Sandoval, C. Gonzalez, R. Alarcon, K. Pichara, and M. Montenegro, "Centralized student performance prediction in large courses based on low-cost variables in an institutional context," *The Internet and Higher Education*, vol. 37, pp. 76–89, 2018.

[20] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," *Computers & Education*, vol. 61, pp. 133–145, 2013.

[21] J. d. J. Costa, F. Bernardini, D. Artigas, and J. Viterbo, "Mining direct acyclic graphs to find frequent substructures—an experimental analysis on educational data," *Information Sciences*, vol. 482, pp. 266–278, 2019.

[22] J. Oliveira, R. Alexandrino, and A. P. Ambrósio, "A survey of applications that use graph-based educational data mining," in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, vol. 30, no. 1, 2019, p. 1401.

[23] K. Danilowicz-Gösele, K. Lerche, J. Meya, and R. Schwager, "Determinants of students' success at university," *Education economics*, vol. 25, no. 5, pp. 513–532, 2017.

[24] E. C. van Rooij, E. P. Jansen, and W. J. van de Grift, "First-year university students' academic success: the importance of academic adjustment," *European Journal of Psychology of Education*, vol. 33, no. 4, pp. 749–767, 2018.

[25] C. Stadtfeld, A. Vörös, T. Elmer, Z. Boda, and I. J. Raabe, "Integration in emerging social networks explains academic failure and success," *Proceedings of the National Academy of Sciences*, vol. 116, no. 3, pp. 792–797, 2019.

[26] R. Arum and J. Roksa, *Academically adrift: Limited learning on college campuses*. University of Chicago Press, 2011.

[27] M. Cachia, S. Lynam, and R. Stock, "Academic success: Is it just about the grades?" *Higher Education Pedagogies*, vol. 3, no. 1, pp. 434–439, 2018. [Online]. Available: https://doi.org/10.1080/23752696.2018.1462096

[28] M. Ben-Ari, "Constructivism in computer science education," *Journal of Computers in Mathematics and Science Teaching*, vol. 20, no. 1, pp. 45–73, 2001.

[29] R. Barros, O. Santana Jr, I. Silva, L. dos Santos, and V. R. C. Neto, "Predição do rendimento dos alunos em lógica de programação com base no desempenho das disciplinas do primeiro período do curso de ciências e tecnologia utilizando métodos de aprendizagem de máquina," in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, vol. 30, no. 1, 2019, p. 1491.

[30] A. P. Carnevale, N. Smith, and M. Melton, "Stem: Science technology engineering mathematics." *Georgetown University Center on Education and the Workforce*, 2011.

[31] R. B. Silva Filho and R. M. de Lima Araújo, "Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis consequências," *Educação Por Escrito*, vol. 8, no. 1, pp. 35–48, 2017.

[32] P. L. Junior, C. Bisinoto, N. S. de Melo, and M. Rabelo, "Taxas longitudinais de retenção e evasão: uma metodologia para estudo da trajetória dos estudantes na educação superior," *Ensaio: Avaliação e Políticas Públicas em Educação*, vol. 27, no. 102, pp. 157–178, 2019.

[33] C. Herman, H. Donelan, H. Jefferis, and J. Hughes, "Entering stem in later life: examining the motivations of adult women studying computing," 2018.

[34] S. Cheryan, A. Master, and A. N. Meltzoff, "Cultural stereotypes as gatekeepers: Increasing girls' interest in computer science and engineering by diversifying stereotypes," *Frontiers in psychology*, vol. 6, p. 49, 2015.

[35] E. S. Weisgram, R. S. Bigler, and L. S. Liben, "Gender, values, and occupational interests among children, adolescents, and adults," *Child Development*, vol. 81, no. 3, pp. 778–796, 2010.

[36] E. S. Weisgram, L. M. Dinella, and M. Fulcher, "The role of masculinity/femininity, values, and occupational value affordances in shaping young men's and women's occupational choices," *Sex roles*, vol. 65, no. 3-4, pp. 243–258, 2011.

[37] J. Clark Blickenstaff*, "Women and science careers: leaky pipeline or gender filter?" *Gender and education*, vol. 17, no. 4, pp. 369–386, 2005.

[38] D. Beede, T. Julian, D. Langdon, G. McKittrick, B. Khan, and M. Doms, "Women in stem: A gender gap to innovation. esa issue brief# 04-11." *US Department of Commerce*, 2011.