Analysis of Student Performance and Social-economic Data in Introductory Computer Science Courses at the University of Brasília

Rodrigo Fonseca Silveira
Depart. of Computer Science
University of Brasilia
Brasilia, Brazil
0000-0003-4965-8519

Maristela Holanda
Depart. of Computer Science
University of Brasília
Brasília, Brazil
0000-0002-0883-2579

Guilherme N. Ramos
Depart. of Computer Science
University of Brasilia
Brasilia, Brazil
0000-0001-5859-7362

Marcio Victorino
Faculty of Information Science
University of Brasilia
Brasilia, Brazil
0000-0003-2785-8958

Dilma Da Silva
Depart. of Computer Science and Engineering
Texas A&M University
College Station, United States
0000-0001-6538-2888

***Abstract*—Computer Science 1 (CS1) courses introduce undergraduate students to computational thinking and their first programming language. As in most institutions, CS1 is a challenge for students at the University of Brasilia, one of the top 10 universities in Brazil. In 2012, the Brazilian higher education system changed with an affirmative-action policy to admit more students from the public K-12 system: the "Quota" Law was implemented at all federal public universities. This paper aims to answer two research questions: 1) What knowledge about the positive/negative impact of certain features on the success of a CS1 course can be discovered from mining educational data augmented by social-economic information? 2) Are these features different between quota and non-quota students? The analysis uses social-economic and academic performance data of undergraduate students from 2012 to 2019. Data mining algorithms such as generalized linear model, gradient boosting machine, and random forest were applied to the data. The findings include: (1) the relevance of indicators such as the consumption rate of university-subsidized meals, (2) that gender is not a determining factor in failure/success, and (3) a higher failure rate for quota students in the Computer Engineering and Mechatronics Engineering majors.**

***Index Terms*—Educational Data Mining; CS1; Introduction to Computer Science course, Machine Learning**

I. Introduction

Introduction to Computer Science courses, commonly called CS1 courses, are a challenging step for many students and an important subject for research in computing education [1]–[7]. Introductory courses have been identified as a pivotal point for recruitment and retention in Computer Science (CS) majors [1]–[4]. These courses, in general, have a high failure rate at institutions around the world, including Brazilian universities [8]. The data used in this work shows a failure rate of over 50% for CS1 at the University of Brasília.

The majority of the top higher education institutions (HEI) in Brazil [9] are public ones and operate free of charge for students. In contrast to the high quality of the education offered by public universities, public high schools fail to prepare students for higher education. Across the country, students from public high schools have a lower performance when compared to those from private ones [10]. Although Brazilian private high schools target the wealthiest members of the population, many families consider them to be the only path to secure admission to the public universities and make financial sacrifices to pay the high cost. The perception is that such investment will pay off if it leads to a high-quality, tuition-free education at a public university.

Traditionally, students secure admission to public universities through entrance exams that cover all elements in the high school curriculum. In 2012, the higher education system in Brazil changed with the introduction of the "Quota" Law [11], an affirmative-action approach to expanding the access of students from public high schools to the federal public university system. Most of the top 10 Brazilian HEIs are funded by the Federal Government. The entrance exam is still used to rank students, but the law reserves 50% of the admission slots for students from public high school. From these quota slots, half of the slots are allocated to students from low-income families (defined as having monthly income of up to one and a half times the national minimum wage), black and indigenous students. Before this law, students from public high schools were a small minority in CS majors in the top HEIs in Brazil [12]. Currently, they represent almost half of such enrollments. In the Department of Computer Science at University of Brasília only 20% of students were from public schools before this law [13]. Clearly, the Quota Law in Brazil impacted diversity in a very positive way, in particular in terms of social-economic profiles.

At the University of Brasilia (UnB), one of the top 10

public universities in Brazil, the Department of Computer Science offers CS1 courses to four majors: Computer Science, Computer Education[1], Computer Engineering and Mechatronics Engineering[2]. These majors exhibit a high drop-out rate, peaking at 64% in the Computer Science and 84% in the Computer Education majors in the years 2013-2018 [14]. In previous studies [15], the CS1 course was identified as one of the most challenging for students.

In the context of first-year students with different educational backgrounds being admitted to majors in the Department of Computer Science, this paper has the following research questions: *1) What knowledge about the positive/negative impact of certain features on the success of a CS1 course can be discovered from mining educational data augmented by social-economic information?* and *2) Are the features different between quota and non-quota students?*. To answer these research questions, we used techniques such as Generalized Linear Model (GLM), Gradient Boosting Machine (GBM) and Random Forest (RF) with social-economic and academic performance data from 2012 (when the Quota law became effective) to 2019. Such exploration aims to inform new initiatives and policies to improve retention in the Department of Computer Science.

The rest of this paper is organized as follows. Section II provides background information on access to Brazilian public higher education and points to some relevant papers on educational data mining. Section III describes the methodology used in our analysis. The results are discussed in Section IV. Section V has the limitations. Section VI presents concluding remarks and future directions.

II. Background

This section presents the system of admission to higher education education in Brazil, as well as a brief contextualization of educational data mining (EDM).

A. Admission to University of Brasilia

Access to public higher education in Brazil depends on the candidates' results in a comprehensive and extremely competitive exam due to the limited capacity available. These exams are designed to test in-depth knowledge in all subjects in the high school curriculum.

The 2018 report Programme for International Student Assessment (PISA) [10] from the OECD (Organisation for Economic Co-operation and Development), which aims to assess student learning outcomes on a global scale, presented the disparity between public and private school K-12 education in Brazil. For Mathematics, Science and Reading Comprehension exams, private schools achieved excellent results, while public schools ranked among the worst in the 79 countries assessed. A detailed study by Esteves and Belluzzo [16] corroborates this wide discrepancy.

[1]A teaching degree for K-12 education.

[2]Mechatronics Engineering blends mechanical and electrical engineering, targeting fields such as robotics and automation.

In an effort to mitigate the bias towards often better-funded private schools and to increase diversity in the public higher educational system, the Brazilian Federal Government approved the Law number 2.711 in 2012. This legislation, known as the "Quota Law", establishes that 50% of new federal HEI students must have completed all of their high school education in a publicly funded school. Furthermore, half of this quota (i.e., 25% of all slots) must be filled by students from low-income families (defined as having a monthly income of up to one and half times the federal minimal wage), or self-declared black or indigenous people [11]. These rules were gradually applied (increasing by 12.5% every year) until 2016, and the results will be reviewed in 2022.

UnB is part of the Brazil's public higher education system. It currently has more than 40,000 undergraduate students, distributed over 150 majors, and almost 10,000 graduate students in 91 Masters and 80 PhD program.

Considering students in the Department of Computer Science in two distinct classes, those from a public school background and those from a private one, Figure 1 shows the trends in admissions from public schools since 2008, four years before the Quota Law. The figure shows the percentages of students from public schools, which clearly increased from around 20% in the years before 2012 to around 50% in 2016, as required by the Quota Law, and its stabilization at this level in the following years.

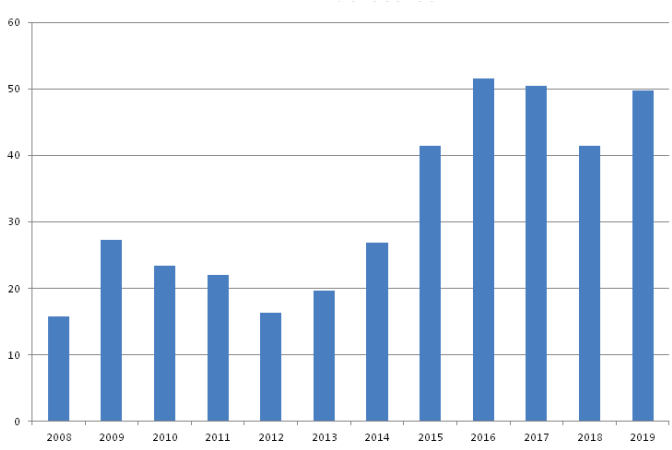

Fig. 1: Percentage of public school students enrolled in CS-related majors at UnB, by year.

However, although the Quota Law is an important tool to encourage social-economic diversity, it only impacts the admission of students, who still have to pursue academic success in their majors. Further supportive measures to address gaps in student preparation or lack of resources were not provisioned in the law. UnB has several support programs, which involve payment for free transportation, free meals, and psychological support, among others. An understanding of these social factors can be useful for departments to better support students, especially during their first year, which has the highest student attrition.

B. Educational Data Mining

Educational Data Mining (EDM) can be defined as the application of data mining techniques to investigate problems in the educational context [17]–[20]. EDM uses educational system databases to derive insights that may lead to the development of educational policies to improve student academic performance. Over the years, several studies have attempted to identify patterns or prediction models for undergraduate student drop-out rates [21]–[27].

Specifically in the context of CS1 courses, many EDM efforts succeeded. Desay et al. [28] deployed social network analysis techniques to the interactions on a CS1 course at a US university; the results revealed that participation in online discussion forums had a positive impact on student grades. Ahadi et al. [29] analyzed the students' performance and consistency in programming assignments in an introductory programming course at the University of Helsinki, Finland using t-Distributed Stochastic Neighbor Embedding (t-SNE) clustering and principal component analysis techniques, which aim to identify students who are at risk of dropping out of their course. Castro-Wunsch et al. [30] investigated early identification of students in need of assistance in computer programming courses. The authors applied neural network-based approaches with Bayesian and decision tree techniques in North American universities.

Many papers [31]–[35] use EDM techniques to analyze students' performance in CS1 courses based on gender and race/ethnicity. In these papers, EDM techniques such as correlations, ANOVA, K-means, and decision tree have been applied to undergraduate students data in US colleges. Kumar [35] considers levels of prior self-confidence and preparedness in the assessment of an online tutor. Trytten and McGovern [31] use correlations of grades and GPA as the predictor of success in CS courses. Quinn et al. [32] report on student success in CS1 courses and persistence. Malla et al. [33] present a prediction model for CS2 outcome (success or failure), based on previous courses such as CS1 and English. Norouzi and Hausen [34] report on an evaluation of student engagement on a CS1 course using an automatic grading system.

As already noted, EDM has been applied to different studies to identify the success of students in computing courses. However, in most of these it is used to analyze grades, online tutors, and interaction between students, which shows the importance of EDM for prediction in educational environments. This paper applies data mining to investigate the experience of CS1 first-year students at the Department of Computer Science of the University of Brasilia. The features that can influence success or failure in CS1 courses are identified. In addition, differences between quota students and non-quota students are investigated. The next sections present the methodology and the findings.

III. Educational Data Mining Process

The steps taken in this research were planned to create a model for predicting outcomes of student performance at the end of a one-year university cycle. Each step is described in the following subsections.

A. Data Understanding

The data collected for this study ranges from 2012 to 2019 and covers four majors in the Department of Computer Science of the University of Brasília, for a total of 2,862 students. These records were then filtered according to the profile of interest, that is, who completed the CS1 course (transfer credits are excluded), resulting in 2,051 anonymous samples. The 21 older students (over 40) were considered outliers and removed. Within this final sample of 2,030 students, 728 are quota students (i.e. admitted through the Quota Law's criteria) and 1,302 are non-quota students.

Table I describes the data used in the experiments, as well as the absolute numbers and percentages by quota and non-quota groups. The features are:

- assistance: if the student received financial support from the university[3];
- previousMajor: if the student transferred from another major;
- age: student's age at the time of enrolment at the university;
- publicSchool: if the student is from a public high school;
- subsidizedMeal: number of meals at the university cafeteria per week[4];
- gender;
- major: the major that the student is enrolled in at Department of Computer Science;
- eveMajor: if the major is an evening offering[5];
- distance: the distance from the student's home to the Campus[6];
- firstRound: if the major was declared as the student's first choice in the admission application;
- success: if the student passed or failed the CS1 course.

Figure 2 shows, by year, the percentage ratios of unsuccessful students in CS1 for both groups, and it is clear that quota students (in blue) are less successful than non-quota ones (in red). As the two groups of students have different failure rates, the factors influencing success in the course for the two groups of students (quota and non-quota students) will be investigated.

B. Data Transformation

To mine the data, it must first be retrieved and then transformed into the state needed for analysis. Information

[3]In the Brazilian public universities, students pay no tuition or fees. Stipends or any kind of financial support are very uncommon due to the limited availability of funding. This help is for undergraduate students in particularly vulnerable social-economic situations.

[4]The university cafeteria offers three meals a day (breakfast, lunch and dinner). In the Brazilian higher education system, federal university cafeterias are subsidized by the federal Government, so the student only pays a nominal amount for food, currently approximately US$1.00 for a full meal. Low-income students can apply for free meals.

[5]Evening majors have all classes starting after 7 pm. They are designed to accommodate students with a full-time job.

[6]In Brazil, most undergraduate students are local residents who live with their families during college.

TABLE I: Data Features

Feature	Statistics			
	Non-Quota		Quota	
assistance:true	20	1.53%	153	21.1%
previosMajor:true	173	13.28%	82	11.26%
age:Between 18 and 20	981	75.34%	553	75.96%
age:Between 21 and 25	104	7.98%	91	12.5%
age:Between 26 and 30	50	3.84%	23	3.16%
age:More than 30	30	2.3%	9	1.23%
age:Up to 17	137	10.52%	52	7.14%
publicSchool:true	166	12.75%	611	84.00%
subsidizedMeal:0	913	70.12%	385	53.00%
subsidizedMeal:Between 1 and 2	286	21.96%	194	26.65%
subsidizedMeal:Between 3 and 5	85	6.53%	100	13.73%
subsidizedMeal:Between 6 and 10	14	1.07%	39	5.35%
subsidizedMeal:More than 10	4	0.03%	10	1.37%
gender:M	1144	87.86%	626	86.00%
gender:F	158	12.14%	102	14.00%
major:ComputerEngineering	357	27.42%	200	27.47%
major:ComputerScience	365	28.03%	229	31.45%
major:ComputerEnducation	220	16.89%	99	13.6%
major:Mechatronics	360	27.65%	200	27.47%
eveMajor:true	220	16.89%	99	13.6%
distance: 11 and 20 km	493	37.86%	282	38.73%
distance: 21 and 30 km	232	17.82%	143	19.64%
distance: 6 and 10 km	244	18.74%	60	8.24%
distance:More than 30 km	67	5.15%	163	22.4%
distance:Up to 5 km	266	20.43%	80	11.00%
firstRound:true	1095	84.10%	623	85.57%
success:true	868	66.66%	326	44.78%

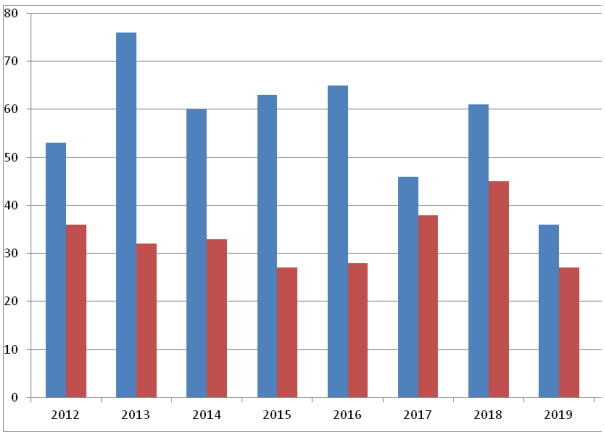

Fig. 2: Failure rate comparison between quota (blue) and non-quota (red) students.

on the CS students was collected using SQL queries from a DBMS PostgreSQL, considering only students since 2012 in one of the four selected degrees who completed the CS1 course (successfully or not) in that same period of enrollment.

Some features were changed from numerical to categorical. The postal codes for the students' home addresses were transformed into an estimated location by a simple calculation of its Euclidean distance to the campus and split into five ranges: under 6 km, 6 to 10 km, 11 to 20 km, 21 to 30 km, and over 30 km. In our dataset, 121 students who had no address information received the average value of 18.1 km. The age (at the time of enrollment) feature was also was classified into five ranges: under 18, 18 to 20, 21 to 25, 26 to 30, and over 30. Finally, the number of meals per week at the university cafeteria were grouped into under 3, 3 to 5, 6 to 10, and over 10.

TABLE II: Models AUC Values - Validation x Test data

Model	Full		Quota		non-Quota	
	Valid.	Test	Valid.	Test	Valid.	Test
GLM	0.8299	0.7377	0.8159	0.7692	0.8551	0.7653
GBM	0.8043	0.7156	0.7744	0.7412	0.8007	0.7307
RDF	0.8258	0.7205	0.7841	0.7342	0.8181	0.7576

C. Data Mining

Once the data is properly structured, the next step is to apply techniques to create the models in order to investigate the structural pattern of classification models for unsuccessful freshmen in the CS1 course. To this end, three supervised learning methods were selected: Generalized Linear Model (GLM), Gradient Boosting Machine (GBM) and Random Forest (RF). The choice of these methods was based on their successful performances evaluating academic data [36]–[41]. All algorithms were applied using the default values for their hyperparameters in the H2O implementaion[7]. The dataset was also balanced automatically. For training the models, the data was split into training (50%), validation (30%), and test (20%) sets.

These procedures were applied to three instances of the dataset: the full set and its two disjoint subsets for quota and non-quota students. Thus, the different models produced could be directly compared using the area under the curve metric (AUC). The results can be observed in Table II. When validating all datasets, GLM produced the best result (over 80%), though only slightly better than the others. Experiments on the test data showed results of around 75%, indicating that the models are more than adequate to the task considering that no academic features were used. As shown in Section II-B, the most successful models heavily rely on academic features.

Further analysis of the GLM model, which provided the better results, yields valuable information since the coefficient values indicate the feature's importance to the result. The values for the model obtained with the full data set are listed in Table III. The results for the same experiments using the quota students' data is given in Table IV, and using the non-quota in Table V. As can be seen, each table has a different set of features, this is because the H2O, used in the experiments, presents as a result only the features that had influence on the result, in this case the student's success in CS1. In these tables, the higher the value of the feature, the greater the influence of the feature on the final result. If the value is positive it means that the feature favors the student's success in CS1, if the value is negative, the feature impairs success in CS1.

IV. Findings and Discussions

This section addresses the proposed research questions.

[7]https://docs.h2o.ai

TABLE III: GLM Full Dataset Coefficients

Coefficient	Value	Signal
subsidizedMeal.Between 6 and 10	1.3742	+
helpRequests.1	0.7684	-
subsidizedMeal.0	0.7529	-
subsidizedMeal.More than 10	0.6371	+
campusDistance.Up to 5	0.3796	+
publicSchool.0	0.3365	+
subsidizedMeal.Between 1 and 2	0.2834	-
campusDistance.More than 30	0.2673	-
publicSchool.1	0.2645	-
racQuota.0	0.2603	+
racQuota.1	0.2060	-
age.Between 18 and 20	0.1877	+
major.Mechatronics	0.1635	+
incomingAge.Between 21 and 25	0.1584	+
major.Computer_Engineering	0.1530	-
age.More than 30	0.1337	+
previosMajor.0	0.1203	+
incomingAge.Between 26 and 30	0.1161	-
firstCall.1	0.1081	+
campusDistance.Between 6 and 10	0.0944	+
major.Computer_Science	0.0944	+
firstCall.0	0.0772	-
gender.M	0.0697	+
previosMajor.1	0.0647	-
gender.F	0.0528	-
subsidizedMeal.Between 3 and 5	0.0519	-
eveMajor.0	0.0389	+
campusDistance.Between 21 and 30	0.0207	-
eveMajor.1	0.0088	-
major.Computer_Education	0.0088	-

TABLE IV: GLM Quota Dataset Coefficients

Coefficient	Value	Signal
incomingAge.Between 26 and 30	0.5187	-
subsidizedMeal.0	0.4859	-
subsidizedMeal.Between 1 and 2	0.4377	-
major.Computer_Engineering	0.4011	-
age.Up to 17	0.3892	+
subsidizedMeal.Between 6 and 10	0.3699	+
eveMajor.1	0.2852	+
major.Computer_Education	0.2852	+
previosMajor.1	0.2753	-
major.Computer_Science	0.2684	+
campusDistance.More than 30	0.2632	-
age.Between 21 and 25	0.2542	-
major.Mechatronics	0.2228	-
subsidizedMeal.Between 3 and 5	0.1942	+
helpRequests.1	0.1025	-
helpRequests.0	0.0965	+
campusDistance.Between 11 and 20	0.0682	+
gender.M	0.0667	-
publicSchool.1	0.0568	-
gender.F	0.0483	+
publicSchool.0	0.0357	+
age.Between 18 and 20	0.0179	+

A. What knowledge about the positive/negative impact of certain features on the success of a CS1 course can be discovered from mining educational data augmented by social-economic information?

The results for the GLM model from the full dataset II(quota and non-quota students) and the values for GLM coefficients (Table III) showed that the frequent use of the university-subsidized meals is an important feature to predict whether a student will succeed or not in CS1 - it is in the top

TABLE V: GLM Non-quota Dataset Coefficients

Coefficient	Value	Signal
subsidizedMeal.Between 6 and 10	2.4534	+
subsidizedMeal.More than 10	1.6683	+
subsidizedMeal.0	1.0859	-
subsidizedMeal.Between 1 and 2	0.9192	-
publicSchool.0	0.4624	+
helpRequests.1	0.4366	-
age.More than 30	0.3830	+
publicSchool.1	0.3280	-
eveMajor.0	0.3112	+
campusDistance.Up to 5	0.2777	+
major.Mechatronics	0.2744	+
major.Computer_Education	0.2262	-
eveMajor.1	0.2262	-
age.Between 21 and 25	0.2191	-
firstCall.1	0.2189	+
subsidizedMeal.Between 3 and 5	0.1706	-
firstCall.0	0.1574	-
previosMajor.1	0.1510	+
major.Computer_Engineering	0.1354	+
age.Between 18 and 20	0.1346	-
campusDistance.Between 21 and 30	0.1272	-
campusDistance.Between 6 and 10	0.1013	+
major.Computer_Science	0.0844	-
gender.M	0.0800	+
age.Up to 17	0.0764	+
gender.F	0.0636	-
campusDistance.Between 11 and 20	0.0030	-

two spots for positive values (*subsidizedMeal.Between 6 and 10 = +1.3742 and subsidizedMeal.Morethan 10 = + 0.6371*). Regarding the university cafeteria at UnB, it is important to highlight some points. As previously mentioned, the cafeteria has a subsidized menu, with the student paying approximately a dollar for each meal, serving three meals a day. Students with vulnerable social conditions can request free meals. During the pandemic period (2020/2021), UnBiversity of Brasilia offered 2,500 free meals to students with verified requirements for food resources [42], [43]. The cafeteria's meals are prepared by a nutritionist, but the menu is limited. Most computing students did not use the university cafeteria, according to the dataset of this study (2012-2019), 64% of students had no meals there. Thus, this feature may be associated with the fact that, among the 36% of students who had meals at the cafeteria, the fact that dining at the cafeteria means they spend more time on campus for their academic agenda and socializing with their fellows.

The feature with the third most positive score involves the distance to the students' home from the university campus. When the students live under 5 km from the campus (*distance.Up to 5 = +0.3796*), they have a higher success rate. There are two possible interpretations of this data. The first is that this is because the main campus is in a middle class neighborhood. Many students who live in these areas belong to an upper social class; unlike in other countries, most students in Brazil continue to live in the parental home throughout their university studies. In the data set, only 17% of students live up to 5 km from the university. The second possible interpretation is that students who live within 5 km of the university save time by having the advantage of a very short commute to the

campus. A large part of the students use public transport (bus) in UnB [44], [45] and for cities that are above 20 km away the journey time can be more than 1 hour. For those who use their own cars, the time in traffic can be long at the busiest time.

Finally, coming from a private school (*publicSchool.0 = +0.796*) and being a non-quota student (*Quota.0=+0.2603*) also stood out as positive features, although in these cases, the value was much lower than that of the previous coefficients. As previously mentioned, students from private schools have a better educational background than those from public schools, and this result was already expected for the first semester courses.

In relation to the negative features, there are, first of all, the students who requested financial aid (*assistance = - 0.7684*) from the university. The university is free of charge, and any student who needs it can apply for a living allowance, a cash sum, or free accommodation and free food. However, this request is granted only for students in vulnerable social-economic situations. In this data set, only 8.5% of the students requested social assistance at UnB. The results show that these students have a high failure rate in the discipline. One action that can be taken in the department is to create support programs to assist students with vulnerability in the computer science department.

B. Are the features different between quota and non-quota students?

To answer RQ2, we divided the experiment into quota and non-quota groups. Quota students had the coefficients with higher values with negative values while for non-quota students, the higher values were positive. In addition, the positive values in the non-quota group have higher values than the positive values of the quota students. This means that, according to the model, non-quota students have more features favorable to success in the CS1 course.

In both groups, quota and non-quota students, the feature of using the university cafeteria had a similar result to the general group. As can be seen for non-quota students, the two positive features are related to frequent meals at the university cafeteria (*subsidizedMeal.Between 6 and 10 = 2.4534* and *subsidizedMeal.More than 10 = 1.6683*), similar to quota students where the second positive feature was also related to the university cafeteria (*subsidizedMeal.Between 6 and 10 = 0.3699*).

The main differences between the quota and non-quota student groups occurred in the following features: for quota students, the Computer Engineering and Mechatronics Engineering majors had negative values, while for non-quota students these values were positive. This result may be related to the need to have a better background in mathematics, so non-quota students perform better in engineering; for the Computer Education major, the quota students had a positive variable, while the non-quota students had a negative value. The Computer Education major is the only night time major in the Department of Computer Science, many students who already have a college degree, or who already work in the field of computing, enter University of Brasilia to obtain their undergraduate degree in the computing area. These students do not enter through the quota system, being part of other types of admission to the university.

With regard to gender, the analysis showed that female quota students had a positive value (*gender.F = +0.0483*), and male quota students a negative value (*gender.M = -0.0667*). This is the opposite when analyzing non-quota students (*gender.M = +0.0800*) and *gender.F = -0.0636*). The rate of girls in computing majors in this data set is 12.81%. The entry of girls into Computing majors at UnB is of approximately 10% in the last 10 years. In a detailed study on the topic, analyzing grades and dropout rates in the CS1 course, it was identified that the social aspects were a higher indicator than the gender feature, which contributes to this result [46]. Regarding the positive point in gender for quota students, the CS Department has the Meninas.comp project [47] which works on the inclusion of girls from public schools in computer courses. The project was created in 2010 and has managed to include girls in different courses at UnB. However, when the trend of failure or success for the group (quota and non-quota) was analyzed, the proportions of the genders, and the small value of the coefficients, it seems that gender cannot be considered a determining factor in success in the CS1 course using this model.

As already mentioned in this paper, the Brazilian educational system is differentiated since public universities, many classified as the best academically in Brazil, are free and have 50% of their undergraduate students entering through the quota system. This change was completed in 2016, and recently a few articles have been published on the topic in the computing area. In the articles [48] and [49], based on a questionnaire survey, results on dropout in computing courses were presented. In [48], similar to the results in the present paper, the dropout rate in the major was higher among students who lived far from the university. In [49] data were presented which showed that low-income quota students have a high dropout rate. Unlike these papers, which are based on other Brazilian universities, here a broader number of variables was used, applied to real data from University of Brasilia's academic and social systems, and this study can facilitate the creation of new policies for student retention in courses. In addition, this paper focuses on the fundamentally important first programming language, so that it is possible to identify the profiles of students who may fail this course.

V. Limitation

Success in the CS1 course for first-year students pursuing a degree in Computing is a challenge that involves different factors. In this paper, the evaluated data was analyzed in search of insights into creating student support policies in the CS Department of University of Brasilia. However, in our analysis, differences in curriculum between the majors, the teaching methodology, the instructor, admission data, among other variables, were not evaluated.

VI. Conclusions

The CS1 course is a challenge for the undergraduate students in the Department of Computer Science at University of Brasilia. To analyze the factors that can influence success in this course, this work presented an analysis of social data and academic performance.

The Quota Law has made a significant contribution to the increase in social diversity in one of the 10 best universities in Brazil. This paper also presented a study on the impact of this change on the failure rate among CS1 students in the Department of Computer Science at the UnB, with analysis of the quota and non-quota groups.

This paper applied EDM to the freshman database from 2012 to 2019. Students are required to take CS1 in their first year. In the analysis, it was noticeable that non-quota students perform better than quota students, and that quota students in the Computer and Mechatronics engineering majors have more difficulty. The university cafeteria is a positive variable for student success in CS1 courses, students who regularly go more frequently to the cafeteria tend to have a better results in the CS1 course. It was also noted that the results were similar between the genders for all quota and non-quota students.

Future works include following the students on their academic trajectory, expanding the analysis to include the subsequent programming course. In addition, EDM could be applied to university admission data, analyzing the grades in the entrance exams for CS majors.

References

[1] C. S. Smith-Orr and A. Garnett, "Motivation and identity in c++ the effects of music in an engineering classroom," in *2016 IEEE Frontiers in Education Conference (FIE)*, (Erie, PA, USA), pp. 1–5, IEEE, 2016.

[2] J. P. Cohoon and L. A. Tychonievich, "Analysis of a cs1 approach for attracting diverse and inexperienced students to computing majors," in *Proceedings of the 42nd ACM Technical Symposium on Computer Science Education*, SIGCSE '11, (New York, NY, USA), p. 165–170, Association for Computing Machinery, 2011.

[3] A. C. Jamieson, L. H. Jamieson, and A. C. Johnson, "Application of non-programming focused treisman-style workshops in introductory computer science," in *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education*, SIGCSE '12, (New York, NY, USA), p. 271–276, Association for Computing Machinery, 2012.

[4] M. S. Kirkpatrick and C. Mayfield, "Evaluating an alternative cs1 for students with prior programming experience," in *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, SIGCSE '17, (New York, NY, USA), p. 333–338, Association for Computing Machinery, 2017.

[5] A. Robins, *The Cambridge Handbook of Computing Education Research*, ch. Novice programmers and introductory programming Anthony V. Robins, pp. 327–376. Cambridge: Cambridge University Press, 2019.

[6] B. A. Becker and K. Quille, "50 years of cs1 at sigcse: A review of the evolution of introductory programming education research," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, SIGCSE '19, (New York, NY, USA), p. 338–344, Association for Computing Machinery, 2019.

[7] R. P. Medeiros, G. L. Ramalho, and T. P. Falcão, "A Systematic Literature Review on Teaching and Learning Introductory Programming in Higher Education," *IEEE Transactions on Education*, vol. 62, no. 2, pp. 77–90, 2019.

[8] C. Watson and F. W. Li, "Failure rates in introductory programming revisited," in *Proceedings of the 2014 conference on Innovation & technology in computer science education*, (Uppsala, Sweden), pp. 39–44, ACM, 2014.

[9] T. W. U. Ranking, "World university ranking 2020." https://www.timeshighereducation.com/world-university-rankings/2020/, 7 2020.

[10] OECD, "Programme for international student assessment (pisa)," tech. rep., OECD, 2018.

[11] Brasil, "Lei Nº 12.711," 8 2012. Accessed April 2022, Available http://portal.mec.gov.br/cotas/docs/lei_12711_29_08_2012.pdf.

[12] S. D. Vasconcelos and E. G. d. Silva, "Acesso à universidade pública através de cotas: uma reflexão a partir da percepção dos alunos de um pré-vestibular inclusivo," *Ensaio: Avaliação e Políticas Públicas em Educação*, vol. 13, pp. 453 – 467, 12 2005.

[13] M. Holanda, M. Mandelli, E. Ishikawa, and D. Silva, "Um relato de experiência do acolhimento d@s calour@s do departamento de ciência da computação da universidade de brasília," in *Anais do XXIX Workshop sobre Educação em Computação*, (Porto Alegre, RS, Brasil), pp. 151–160, SBC, 2021.

[14] CPA/UnB - Comissão Própria de Avaliação da UnB, "Relatório do perfil dos estudantes (2017-2019)," 2020. Disponível em http://cpa.unb.br/index.php?option=com_content&view=article&id=456&Itemid=305, Acessado: Junho, 2020.

[15] M. Holanda, M. Dantas, G. Couto, J. Correa, A. P. de Araújo, and M. E. Walter, "Perfil das alunas no departamento de computação da universidade de brasília," in *Anais do XI Women in Information Technology*, (Porto Alegre, RS, Brasil), SBC, 2017.

[16] A. G. E. d. Moraes and W. Belluzzo, "O diferencial de desempenho escolar entre escolas públicas e privadas no brasil," *Nova economia*, vol. 24, no. 2, pp. 409–430, 2014.

[17] R. S. Baker and P. S. Inventado, *Educational Data Mining and Learning Analytics*, pp. 61–75. New York, NY: Springer New York, 2014.

[18] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017.

[19] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, no. 1, pp. 135 – 146, 2007.

[20] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4, Part 1, pp. 1432 – 1462, 2014.

[21] M. Agaoglu, "Predicting instructor performance using data mining techniques in higher education," *IEEE Access*, vol. 4, pp. 2379–2387, 2016.

[22] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," *Journal of Medical Systems*, vol. 43, p. 162, Apr 2019.

[23] S. M. Merchan Rubiano and J. A. Duarte Garcia, "Formulation of a predictive model for academic performance based on students' academic and demographic data," in *2015 IEEE Frontiers in Education Conference (FIE)*, (El Paso, TX, USA), pp. 1–7, IEEE, 2015.

[24] M. Zaffar, M. A. Hashmani, K. Savita, and S. S. H. Rizvi, "A study of feature selection algorithms for predicting students academic performance," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 5, pp. 541 – 549, 2018.

[25] L. Zhang and H. Rangwala, "Early identification of at-risk students using iterative logistic regression," in *Artificial Intelligence in Education* (C. Penstein Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, and B. du Boulay, eds.), (Cham), pp. 613–626, Springer International Publishing, 2018.

[26] L. C. B. Martins, R. N. Carvalho, R. S. Carvalho, M. C. Victorino, and M. Holanda, "Early prediction of college attrition using data mining," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1075–1078, IEEE, 2017.

[27] R. da Fonseca Silveira, M. Holanda, M. de Carvalho Victorino, and M. Ladeira, "Educational data mining: Analysis of drop out of engineering majors at the unb-brazil," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 259–262, IEEE, 2019.

[28] U. Desai, V. Ramasamy, and J. D. Kiper, "A study on student performance evaluation using discussion board networks," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pp. 500–506, 2020.

[29] A. Ahadi, R. Lister, S. Lal, J. Leinonen, and A. Hellas, "Performance and consistency in learning to program," in *Proceedings of the Nineteenth Australasian Computing Education Conference*, pp. 11–16, 2017.

[30] K. Castro-Wunsch, A. Ahadi, and A. Petersen, "Evaluating neural networks as a method for identifying students in need of assistance,"

in *Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education*, pp. 111–116, 2017.

[31] D. A. Trytten and A. McGovern, "Moving from managing enrollment to predicting student success," in *2017 IEEE Frontiers in Education Conference (FIE)*, pp. 1–9, IEEE, 2017.

[32] B. A. Quinn, W. M. DuBow, and D. Sul, "Understanding who enrolls in introductory computing courses at community colleges," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, pp. 49–55, 2019.

[33] S. Malla, J. Wang, W. Hendrix, and K. Christensen, "Predicting success for computer science students in cs2 using grades in previous courses," in *2019 IEEE Frontiers in Education Conference (FIE)*, pp. 1–5, IEEE, 2019.

[34] N. Norouzi and R. Hausen, "Quantitative evaluation of student engagement in a large-scale introduction to programming course using a cloud-based automatic grading system," in *2018 IEEE Frontiers in Education Conference (FIE)*, pp. 1–5, IEEE, 2018.

[35] A. N. Kumar, "Results from repeated evaluation of an online tutor on introductory computer science," in *2011 Frontiers in Education Conference (FIE)*, pp. T2H–1, IEEE, 2011.

[36] V. L. Miguéis, A. Freitas, P. J. Garcia, and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," *Decision Support Systems*, vol. 115, pp. 36–51, 2018.

[37] 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.

[38] P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira, *et al.*, "A comparative study of classification and regression algorithms for modelling students' academic performance.," *International Educational Data Mining Society*, 2015.

[39] A.-S. Hoffait and M. Schyns, "Early detection of university students with potential difficulties," *Decision Support Systems*, vol. 101, pp. 1–11, 2017.

[40] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura, "Web usage mining for predicting final marks of students that use moodle courses," *Computer Applications in Engineering Education*, vol. 21, no. 1, pp. 135–146, 2013.

[41] J.-P. Vandamme, N. Meskens, J.-F. Superby, *et al.*, "Predicting academic performance by data mining methods," *Education Economics*, vol. 15, no. 4, p. 405, 2007.

[42] J. S. Lima, "A assistência estudantil na universidade de brasília durante a pandemia do covid-19," *Cadernos Cajuína*, vol. 6, no. 3, pp. 228–242, 2021.

[43] M. C. de Sant'Anna and G. E. Moreira, "Caracterização da política de assistência estudantil: Um enfoque na universidade de brasília," *Revista Ciências Humanas*, vol. 12, no. 3, 2019.

[44] A. O. Aruwajoye, "Previsão de demanda de transporte no campus darcy ribeiro da universidade de brasilia," tech. rep., Universidade de Brasília, 2016.

[45] M. d. L. Matos, "Relatório técnico de proposta de reestruturação da infraestrutura da universidade de brasília baseada no conceito de smart campus," tech. rep., Universidade de Brasília, 2021.

[46] K. Ketulhe, M. Holanda, A. Lima, A. Borges, A. Araujo, C. Castanho, C. Koike, and R. Oliveira, "Análise do desempenho acadêmico das alunas cotistas na primeira disciplina de programação da universidade de brasília," in *Anais do XVI Women in Information Technology*, (Porto Alegre, RS, Brasil), pp. 1–11, SBC, 2022.

[47] M. Holanda, A. P. Araujo, and M. E. Walter, "Meninas.comp project: Programming for girls in high school in brazil," in *2020 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*, vol. 1, pp. 1–2, IEEE, 2020.

[48] T. Rocha, E. Santos, V. Júnior, and C. Souza, "Comparação entre o perfil dos evadidos e dos egressos de um curso de tecnologia," in *Anais do XXVII Workshop sobre Educação em Computação*, (Porto Alegre, RS, Brasil), pp. 404–413, SBC, 2019.

[49] M. A. C. Pena, D. A. S. Matos, and R. M. d. E. Coutrim, "Percurso de estudantes cotistas: ingresso, permanência e oportunidades no ensino superior," *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, vol. 25, pp. 27–51, 2020.