The Main Factors of Enem: A Literature and Microdata Perspective

1st Jacinto José Franco
Computer Science
IFMT
Barra do Garças-MT, Brazil
jacinto.franco@ifmt.edu.br

2nd Fernanda Luzia de A. Miranda
Teaching
IFMT
Barra do Garças-MT, Brazil
fernanda.miranda@ifmt.edu.br

3nd Jacques Duilio Brancher
Computer Science
Londrina State University
Londrina, Brazil
jacques@uel.br

4th Lirian Keli dos Santos
Teaching
IFMT
Barra do Garças-MT, Brazil
lirian.santos@ifmt.edu.br

***Abstract*—This is a Research Full Paper. Enem is a test applied annually since 1998 in Brazil with the purpose of evaluating high school students. The data from the tests are made available in Comma-separated values (CSV) files by Inep, enabling analysis to be carried out on more than 102 million records, combining factors and levels with a wealth of detail from a social and educational point of view. This is one of the most consistent databases in the world and yet there is no article that aims to establish, from the literature, the most frequent factors, relating the factors evidenced with data mining to interactions. From this discovery, it was possible to create unique visualizations of the last 25 years, exploring interactions between factors and levels, which enabled the discovery of new knowledge. To guide this study, a bibliographic search was carried out of all the articles returned in the Google Scholar search, making it possible to identify the most referenced factors. With this information, data was pre-processed using the Pareto principle (80/20), calculating averages and establishing grade levels and income levels in order to normalize the data, since the tests have varied in difficulty over the years and there have been significant changes in the economy since the first version of the exam. As for student performance, the financial investments made at primary and secondary level are enough to provide an indication to establish assumptions and, for this reason, family income is strongly related to educational opportunities and all the main factors. The best primary and secondary education institutions are federal and private schools. Skin color exerts some influence, but does not interfere significantly when one has advantageous financial conditions. The results highlight the importance of thinking about Brazil's poorest population, because as well as being socially vulnerable and disadvantaged, their performance is below what can be achieved.**

***Index Terms*—Enem microdata, factors, student performance, visualization, literature perspective, feature engineering**

I. Introduction

The Brazilian government has administered the National High School Exam (Enem) annually since 1998 to assess the knowledge and skills of high school students [1]. The exam is conducted by the Anísio Teixeira National Institute for Educational Studies and Research (INEP), which is linked to the Ministry of Education (MEC).

Because of its scope, universities have gradually used the test as the main objective criterion for the selection of Brazilian students in public and private universities [2], and it is now widely accepted. All exams and anonymized results are released regularly, making it possible to analyze the results, factors and interactions in more than 102 million tests. As a result, this is the most structured and comprehensive set of educational data on access to higher education in Brazil and one of the largest in the world [3], second only to China's Gaokao, which provides a high capacity for generalization about secondary education. Despite this, there is still no academic work that analyses the entire database of more than 102 million Enem tests using visualization resources for these elements.

In this sense, this article seeks to provide a global vision when thinking about the performance of high school students, based on data collected between 1998 and 2022. The study is guided by the main factors identified in the literature on the statistical analysis of the Enem databases. It also relates to the most important factors already ranked in [4] work, which applied classifiers and selectors. In this way, the interactions between the factors led to the generation of unique visualizations of the 25 years of Enem assessments.

To achieve this, this work makes use of systematic review tools to extract the main factors from the literature, SQL (Structured Query Language) to summarize the most important information and Python language packages to visualize the data over the years. In addition, it is important to say that, in order to facilitate analysis, all the evidence was pre-processed, based on the 80/20 Pareto concept for income [5].

For the scores, the engineering of the characteristics followed what was initially established by [4], thus providing a normalization of data, due to the differences in both the level of difficulty of the test and the significant changes in society over the last 25 years. In addition, using binary levels reduces the complexity of the levels and makes it easier to analyze the information.

The article is justified because there is no study in the literature that explores all of Enem's bases and, therefore, there is a need to understand the educational evolution of the last two decades, contributing to discussions on educational policies and concrete interventions for quality education.

In terms of organization, in addition to the introduction (section 1), this article presents section 2, which sets out the detailed methodological design. Section 3 establishes the main factors extracted from the literature and presents a discussion of them. Section 4 visually shows the information on the factors, levels and their interactions and, finally, section 5 presents the conclusions.

II. Methodology

This work makes use of systematic literature review tools in order to summarize the main factors referenced in the literature, which makes it possible to relate them to the factors ranked by [4] using classifiers and attribute selectors. The systematic review was chosen because it is an important methodological resource capable of providing consistent answers on the literature [6] and, together with more than 102 million Enem data points, it provides a broad evidence base on the performance of Brazilian high school students.

The review was structured with the intention of answering the following question: "What are the main factors in the articles already published on Enem datasets?" The search engine used in the research was Google Scholar, to identify articles that fit the established protocol. Only scientific articles that did not apply data mining resources were considered, since the work of [4] already provided initial high-level evidence on the factors, but did not explore the possible combinations with visualization.

Once we had the most relevant factors from a literature perspective, the next step was to import all the Enem microdata into a PostgreSQL database. This allowed the data to be pre-processed and extracted using SQL to generate graphs and textual analysis. This provides a more complete overview of the most relevant and frequent factors.

In the pre-processing stage, the Pareto concept [5], 80/20, was used to categorize the students according to their social group, thus making it possible to analyze the microdata over the years in relation to income. This is due to the considerable changes in society over the last 25 years, making a direct analysis naive and inaccurate. Therefore, a comparison was made between the bottom 80% of socio-economic groups and the top 20%, as these disparities have a significant impact on access to quality education.

When it came to the scores, all the averages were calculated and a label "Above" was created for the scores that were 14% above the average. This process was replicated for all Enem years. The incorporation of these stratifications in score and income allows the suppression of insignificant intermediate categories [7].

Finally, the work focuses on analyzing the main factors graphically and through tables in order to establish combinations of factors and unique levels not yet available in the specialized literature, thus providing a greater understanding of the student performance of Brazilians over the last two and a half decades. The analyses focused on establishing percentages of "Above" grades and the number of students required to be labeled as "Above", as this makes the analyses simple enough to understand the performance of millions of tests over the last 25 years.

III. The most common and most important factors

This section establishes the most referenced factors in order to relate them to those stipulated by [4], providing an overview of more than 102 million pieces of data, generating interactions between the main factors, both from the point of view of data mining and the literature on the subject.

In order to guide the analysis undertaken in this work, a search was carried out on Google Scholar with the search string "Enem microdados". The site returned a total of 1,476 unique entries of academic work, but only 42 of these were considered pertinent to the purposes of this production, as they met the following criteria: being a scientific article, having statistically analyzed the microdata provided by Inep and providing some visualization with tables or graphs of student performance.

For a factor to be mentioned in Table 1, it needs to be mentioned in the summary or conclusions, indicating its relevance.

TABLE I
MAIN FACTORS REFERENCED IN THE LITERATURE ON STUDENT PERFORMANCE IN THE ENEM OF ALL YEARS ON GOOGLE SCHOLAR

Factor	Total
Type of school - high school (public/private/independent)	27
Family income	23
Color / ethnicity	13
Father's or mother's education	12
Gender	12
Locality or region (states)	11
Age	9
Rural or urban area	5

Based on the four main factors shown in Table 1, it became possible to search the microdata in order to base the analysis on what is most important from a literature perspective.

In line with what is shown in Table 1, this study also analyzes the Elementary School and Foreign Language factors. These factors are the most informative according to [4] and,

in addition, it was found that they have not yet received due attention in the literature, which will be explored in this work.

The following sections outline the main interactions [8] of the most important factors with income from a Pareto perspective, which provides a global view of the last 25 years, based on millions of educational data.

IV. Data analysis and knowledge discovery

This section aims to highlight the knowledge learned in the literature and relate it to what can be obtained from the microdata of the last 25 years of Enem, with the analyses being oriented according to the order of the most mentioned factors, as per the results obtained in the review summarized in Table I.

The database consists of 102,968,591 tests, which allows us to identify some patterns in the population. The assessment saw the highest take-up in 2014, as shown in Figure 1. Regarding the total number of registrations, it is known that the time series is highly correlated with the dollarized Brazilian Gross Domestic Product (GDP)[1], with a Pearson correlation factor of 0.80, which gives us a strong indication that the population's purchasing power influences the decision to take Enem or not.

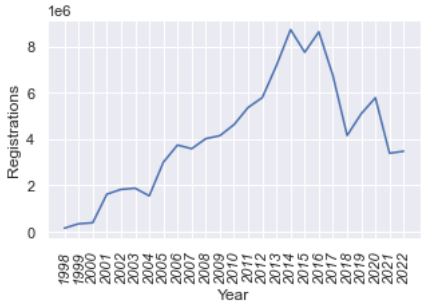


Fig. 1. total number of Enem exams over the last 25 years in millions.

Knowing that the number of registrations is quite significant, it can be said that the visualizations and knowledge gained from these databases are more interesting from a statistical point of view than a lot of the information that has been published in the literature on education. For this reason, the following will be an analysis of the factors most frequently referenced in the literature, including: type of school, family income, color/ethnicity and parents' schooling. In conjunction with these factors, we will analyze primary education, foreign language and family income, as evidenced by [4].

A. Pre-university education

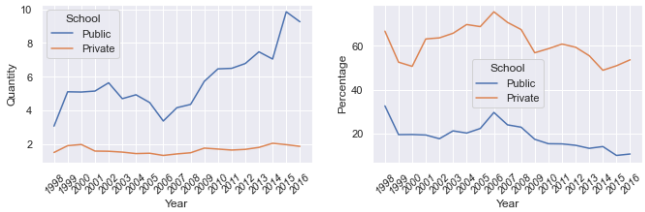
In the microdata, there is information stored on schools at secondary level in all years and primary level up to 2016. However, the information on primary education is inconclusive about the current state of education, as it hasn't been collected for 7 years. However, it still provides an indication of performance.

At primary and secondary school level and in the other factors discussed in the following sections, the focus was on identifying the ratios and percentages that would make it possible to identify the prevalence rates over the years.

In view of the above, this section presents, in an unprecedented way, some of the interactions that can be observed between primary and secondary education in the performance of Brazilian students, as well as a unique view of all the years at secondary level.

1) Elementary School: This subsection aims to highlight student performance based on information about elementary school. This factor deserves a prominent place, as it is the most relevant factor in [4] study and is not among the most studied, according to Table I.

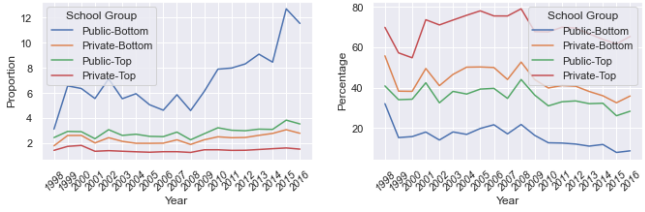
Brazilian elementary school consists of 9 years of study and represents 3/4 of all the initial years of pre-university schooling.



(a) for every number of people there is one "Above" (b) percentage of people Above

Fig. 2. Elementary school performance - people labeled "Above".

From what is illustrated in Figure 2(a) and 2(b), between 1998 and 2016, it became more difficult for the majority of the Brazilian population to obtain a good performance and, consequently, access to courses and universities, especially since 2008.



(a) for every number of people there is one "Above" (b) percentage of people Above

Fig. 3. Elementary school performance - by the financial group (Top and Bottom).

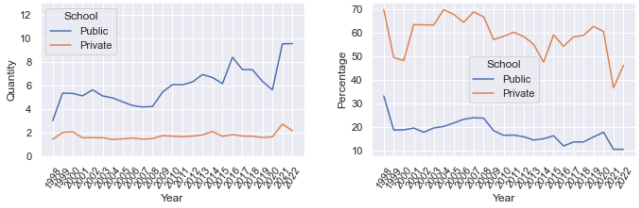
When separating by income, the result is even worse for those with few financial resources. Even in private institutions, the result is better for those at the top financially (Figure 3(a) and 3(b)), and the same is true in public institutions. Although

[1]https://www.worldbank.org/en/home

Enem has not collected elementary school data since 2016, it can be said that there has been no significant change in the public sphere, as the following sections corroborate.

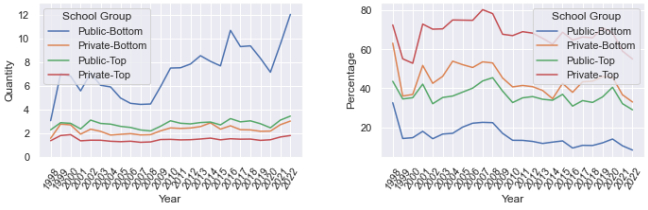
Another point to note is that the vast majority of students who did their primary schooling in private schools continue in the same type of institution in secondary school, but in higher education, these students prefer public universities.

2) High school: According to the literature, secondary education is the most cited factor. For this reason, this section seeks to provide an overview of performance over the years based on the labeled data, making it possible to highlight the performance of the best students.



(a) for every number of people there is one "Above" (b) percentage of people Above

Fig. 4. High school performance

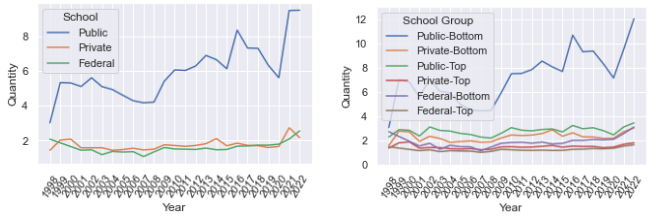


(a) For every number of people there is one "Above" high school by financial group (Top and Bottom) (b) The percentage of people performing "Above" in high school by financial group (Top and Bottom)

Fig. 5. High school performance by the financial group.

Figure 4 (a and b) as well as Figure 5 (a and b) follow the same pattern observed in section IV-A1, on primary education, in which there is a significant worsening in the performance of the poorest students from 2008 onwards, especially for the population of Brazil's financial base, most of whom only have access to public education. Of all the possible combinations, the performance of students from the financial Top has remained the same over the last 25 years and was not so far off when these students attended public schools.

In the public school system in 2022, as filled in by the students, 89% of the enrolments are from the financial base and in the private school system, only 40%. In that year and in the same financial group, 1 in 12 people were labeled as "Above" in the public network, compared to 1 in 3 in the private network. At the financial top, 1 out of every 3.4 people in the public network and, in the private network, 1 out of every 1.8.

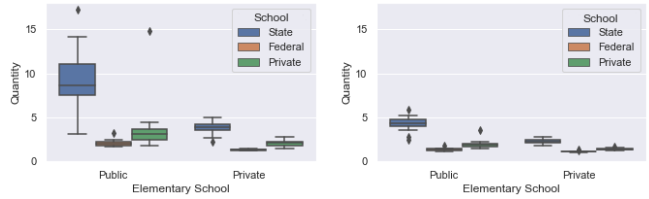


(a) For every one person there is an "Above" in high school (b) The percentage of people performing "Above" in high school by financial group (Top and Bottom)

Fig. 6. High school education performance by the financial group including the federal education system.

The exception to Brazil's public schools are federal schools, as can be seen in Figure 6(a), where the result is slightly better than private schools. Figure 6(b) shows that the best education in Brazil is provided by federal schools when students are better off.

In 2022, 20.6% of all duly identified enrollments are from private schools, while federal public schools account for 5.81% of public enrollments. From a quantitative point of view, there are 4.5 times more students in private schools than in federal schools. Therefore, for the vast majority of Brazilians, any inclusion policies applied in federal schools are inaccessible.



(a) for every number of people there is one "Above" in the Bottom group (b) for every few people there is one "Above" in the Top group

Fig. 7. Combined performance of high school and elementary school by financial group (Bottom and Top). Period: 1998-2016, except 1999 and 2000.

Still on the subject of federal schools, it can be seen that the people who benefit most from the federal network are the students at the financial top who come from private elementary schools; this is also the best combination in elementary and high school. However, federal education is where the gap between the bottom and the top is smallest, regardless of origin, as illustrated in Figure 7(a) and (b).

The worst interaction is when you have few financial resources, your elementary school was in public schools and your high school was in state schools. This is the profile of the vast majority of Brazilians. This is due to the fact that Brazilians in state schools come from worse social and especially financial backgrounds, according to [9] and [10].

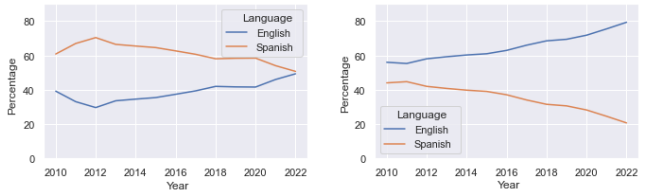
In addition, the teaching staff at private institutions are better than those at public ones. The exception to this premise is the federal institutions, whose investment is compatible with

or better than that provided in the private network and also provide better infrastructure and decent salaries for teachers, thus allowing them to attract more qualified and technically competent teachers.

Regarding the total number of students, based on the 2022 microdata, it is estimated that around 53.3% of the best students in Brazil come from the private school system. Therefore, the private network competes on an equal footing with the entire public education network in Brazil, even though the latter has almost 3.9 times as many students as the private network.

B. Foreign language

At Enem, Brazilian candidates must choose a foreign language when registering, either English or Spanish. Regarding the language, it can be seen that candidates with better financial conditions choose English, as shown in Figure 8(b). There is also a growing preference for English among poorer Brazilians, as shown in Figure 8(a).



(a) percentage of the financial Bottom (b) percentage of the financial Top

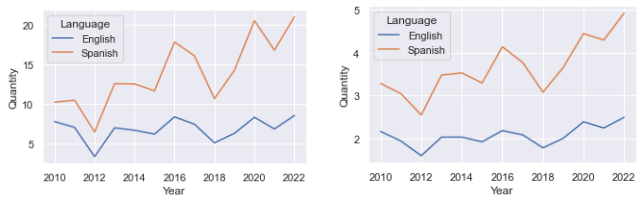
Fig. 8. Foreign language choice of the financial group(Base and Top).

Considering only the students labeled ”Above”, it can be seen that the preference for English is even more pronounced, both at the bottom (Figure 9(a)) and at the financial top (Figure 9(b)). Interestingly, for the bottom (Figure 9(a)) English was more important in the same year that Brazil had the highest GDP peak on record.

(a) percentage of the financial Bottom (b) percentage of the financial Top

Fig. 9. Choice of foreign language over the years by the financial group and labeled ”Above”. Period: 2010-2022.

As shown in Figure 10(a) and 10(b), the financial Top and Bottom reveal that students who choose English are considerably better than students who choose Spanish. In 2022, those who chose English were 3.2 times better than Spanish,



(a) financial Bottom (b) financial Top

Fig. 10. Each one is labeled ”Above” by the language of the financial group. Period: 2010-2022.

for the Bottom the difference is 2.9 times and at the Top only 1.7.

C. Family income

This is one of the factors that most influence the student performance of Brazilians and is the second most referenced in the literature. However, in the work by [4], which ranks the top 10 factors for all years from a machine learning perspective, it can be seen that all of the top 10 factors are strongly influenced by family purchasing power. It is in third place among all the aspects considered about students.

In both views, income is not the main factor, because it’s not just about having money, but how it is used to provide better study and socio-cultural conditions. Income influences where students live and study and the environments and experiences they have access to.

In view of this, this subsection will show some of the evidence that will support the statements about the population, based on the 2022 tests and, subsequently, all the other years.

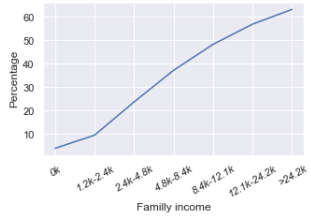
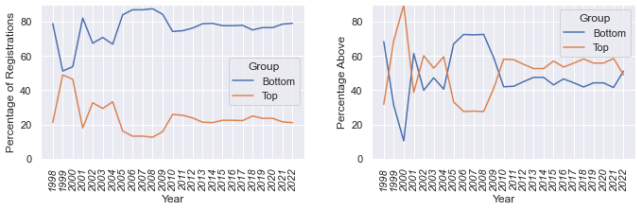


Fig. 11. Percentage of enrollments labeled Above by family income in 2022. Values in BRL.

The average test score and income are variables with a Pearson correlation of 0.75, which means that the probability of someone being labeled ”Above” increases significantly as social conditions improve, as shown in Figure 11.

In 2022, only 19.5% of all Enem exams were labeled ”Above”. In that year, 79% of families earn up to BRL 4,848.00 and have 51% of the best results. The 21% of families who are financially better off got 49% of the excellent results.

Analyzing all the years based on what was modeled and shown in Figure 12 (a), it can be seen that the separation of the top and bottom groups is not perfect, as this is an imposition of the micro-data, and there are inconsistencies in percentage



(a) How much each group has represented over the years

(b) Percentage of registrations labeled as Above by the group

Fig. 12. 80/20 Pareto modeling by income

terms. However, the characterization of the performance of people at the top of the financial ladder has remained almost unchanged over the years, as shown in Figures 3, 4 and 5.

In the last decade, all the people in the financial base did not manage to get the best marks in percentage terms, as can be seen in Figure 12(b) and in the other years, according to Figures 5(a) and 6(b).

(a) Quantity of Top and Bottom to have one Above

(b) Difference between Top and Bottom

Fig. 13. Every few there is one labeled "Above".

As shown in Figure 13(a), those at the Top financially have performed steadily over the last 25 years, maintaining their leadership on the national scene. Therefore, the difference between the Top and the Bottom, as shown in Figure 13(b), has become even greater over the years. Over the last 12 years, the financial Top has accounted for 23% of the tests and holds an average of 55.3% of the best results.

The result in Figure 13(a) is very similar to what was found about the public versus private network in Figure 4(a) and 6(a), as people who are better off generally enroll their children in private schools.

D. Skin color or Race

Color is a factor that has been collected every year in the data sets and is the third most referenced factor in the literature. This is due to the fact that the Brazilian people are made up of multiple peoples and ethnicities. However, this factor is not among the 10 relevant factors to define student performance [4] in the 22 years analyzed.

For this reason, this work defends the idea that identifying where the student went to primary and secondary school, knowing their family income and the choice of foreign language are key elements in suggesting something about student performance and, although there are some subtle differences, it can be seen that what has the greatest impact is the financial group to which the student belongs.

While defending the idea that the color or ethnicity factor alone does not define student performance, this section analyzes the nuances and findings in the data sets. In some studies, it can be seen that more emphasis is placed on the inclusion of mostly black and brown students. However, it turns out that the most excluded group in Brazil are indigenous people, with only 6.6% of their tests labeled as "Above", as shown in Table 2.

TABLE II
SKIN COLOR OR RACE AT ENEM 2022

Skin color/race	Represents	Above
Undeclared	1.74%	23.2%
White	43.77%	34.8%
Black	10.91%	13.5%
Mixed	41.23%	16.9%
Asians	1.87%	24.4%
Indigenous	0.48%	6.6%

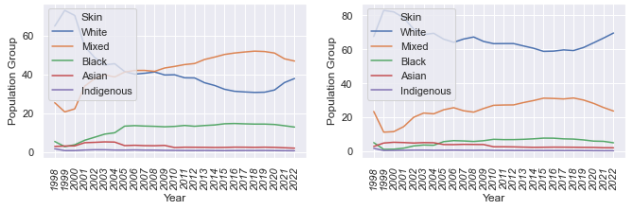
Separating color by financial group, Top and Bottom, we can see that for every color or race, being in a family with better conditions makes all the difference. In this sense, it can be said that a black or brown person who has good financial conditions is better off than 78.5% of the white group in the Base. The same is true for yellow people, but not for indigenous people, as it took 29.8 indigenous people in 2022 for one to be labeled as "Above" in the base performance, as shown in Figure 11 and Table 2.

By separating the top and bottom groups, we can see that the most relevant groups in the Enem for all the years are brown, white and black, in that order, in the bottom group, as shown in Figure 14(a). In the financial top group, the most significant group are whites, browns and blacks, as shown in Figure 14(b). There is therefore relatively little change in the percentage of browns and whites at the top.

From a social point of view, there have been no significant changes in family income over the last few decades, which implies that racial and social quota policies have not brought about significant changes in society. To give you an idea, in 2022, black and brown people represent 52.1% of all registrations, but make up only 28.4% of the top and 59.7% of the bottom. Whites represent 43.77% of registrations, but hold 69.6% of the Top.

From the point of view of performance, at the top (Figure 15(b)) the differences are irrelevant between the color and race categories compared to the bottom (Figure 15(a)). It can be seen that whites perform best, while indigenous people perform considerably worse than whites, browns and blacks.

In 2022, to give you an idea, of the base, one in every 5.4 white people were labeled as "Above", in the same year, one



(a) percentage of Bottom (b) percentage of the Top

Fig. 14. Representation by skin color over the years.

(a) Bottom (b) Top

Fig. 15. Every few people are labeled "Above" because of their skin color or ethnicity.

in every 10.5 brown people were labeled as "Above".

The same information, from a percentage point of view, can be seen in Table 2. It can be seen that the likelihood of someone being labeled "Above" when they are from the social base has a considerable impact, regardless of the color and race category to which they belong. A mixed-race person is much closer to a poor white person than to a mixed-race person at the top. It is clear that the poorest group in Brazil are indigenous people, followed by blacks, browns and whites.

TABLE III
PERFORMANCE VERSUS SKIN COLOR/RACE AT ENEM 2022.

Skin color/race	Group	Percentage of group	Above
Undeclared	Bottom	87,4%	14,6%
	Top	12,6%	**38,4%**
White	Bottom	**78,5%**	25,0%
	Top	21,5%	**45,0%**
Black	Bottom	**93,2%**	8,5%
	Top	6,8%	**29,4%**
Mixed	Bottom	**91,4%**	11,0%
	Top	8,6%	**34,5%**
Asians	Bottom	86,6%	16,2%
	Top	13,4%	**48,8%**
Indigenous	Bottom	95,6%	3,7%
	Top	4,4%	**19,4%**

Table 3 shows that the best performers are yellow, white and brown people. We can also list the poorest group, which is made up of indigenous people, blacks, browns, yellows and whites. Browns and whites are numerous in the poorest group, as they account for 85% of the tests in 2022, and the result is similar in the other years.

Regarding the interactions observed, it can be said that economic capital removes any possible inequality between the color and race categories, but the social interactions in which individuals are exposed generate inequality in all incomes and this social dynamic is not collected by the data sets provided by Inep. What can be said is that there is a color/race category with a higher concentration of poor people, which is why the result is different.

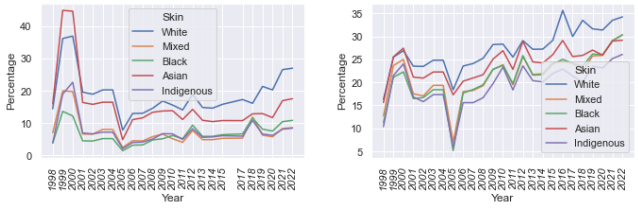
Another combination explored was sex and color/race, in which it can be seen that the majority of registrations are from women, 51.26% of brown and white women, 32.65% of brown and white men. In this sense, it can be seen that the majority group are women and for every man there were 1.6 women enrolled in Enem in 2022. According to the IBGE, Brazil is made up of 48.9% men and 51.1% women, i.e. for every man there are 1.04 women, thus showing that there is inequality between the sexes in terms of access to Enem.

Complementing the information already listed on gender, it is important to note that in 2019, 11 times more men than women died in all homicides. Of the men, 76.72% of homicides were of black men and 21.63% of white men. In the same vein, it can be seen that if we only consider male suicide, the total is higher than all female homicides and suicides [11].

Finally, it can be said that black men die more from homicide and suicide, are poorer and that the probability of being labeled "Above" in performance is relatively low in all years. What reduces or mitigates inequalities is having a higher income, as can be seen in Figure 11.

E. Fathers' schooling

With regard to the education of Brazilian fathers, it can be seen that, from the point of view of performance, the higher the education level of the father of the Enem participant, the higher the score. The effect is more positive than that observed for the educational level of mothers. For this reason, an evolution will be generated over the years, focusing only on the parents' education according to the color or race of those enrolled and their performance.

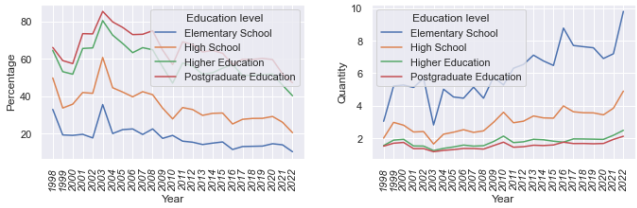


(a) Sum of undergraduate and graduate degrees (b) With high school

Fig. 16. Percentage of fathers' schooling by skin color of registrant.

Figure 16 shows that the parents of white registrants are the ones with the highest levels of education, whether at secondary level (Figure 16(b)) or at undergraduate level (Figure 16(a)).

This partly explains the results for whites shown in the previous sections and subsections.

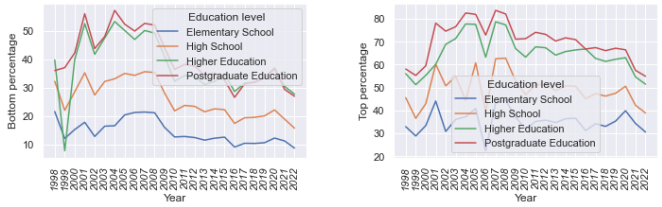


(a) Percentage of students labeled "Above" (b) Every few there is one labeled "Above"

Fig. 17. Father's schooling when students are labeled as "Above".

Figure 17 provides some useful information that allows us to identify how difficult it is for a student to be labeled "Above". In Figure 17(a), when the father has at most primary education, the probability of a student being labeled "Above" is only 10% in 2022.

The probability increases with the father's schooling. Another reasoning that can be made in relation to the data is that, as can be seen in Figure 17(b), a child whose father has only secondary education is twice as likely to be labeled as "Above" when the father has higher education. When the father has at most primary education, it is twice as difficult if he has secondary education.



(a) Bottom (b) Top

Fig. 18. Percentage of registrations labeled "Above" by father's schooling by income.

Evaluating the percentages, it can be seen that the probability of the student being labeled as "Above" increases considerably when the father's schooling is high (Figure 18(a)) and doubles the probability of the application being labeled as "Above" when the income is high (Figure 18(b)).

Finally, it can be seen that the differences in education between fathers with undergraduate and postgraduate degrees are minimal (Figure 19(a)), while the difference is greater between those at the bottom and at the top of the financial ladder, and this discrepancy is accentuated when the father has no more than primary education.

V. Conclusions

From the review process, it was possible to identify the main factors referenced in the literature by reading the abstracts and

(a) Bottom (b) Top

Fig. 19. For every number of registrations, there is one "Above", based on the father's schooling and income. Period: 1998-2022.

conclusions of the articles, providing a view over the years of the four main factors: where the student went to high school, family income, color/race and father's/mother's education. With the Pareto method, separating the bottom versus top groups showed that social position is a more accurate measure of student performance than analyzing the factor in isolation.

Color or race is among the main factors in the literature because Brazil is made up mostly of browns and whites, the result of a miscegenation of different races. However, according to [4], using machine learning algorithms, it was identified that in none of the years analyzed did color or race enter the top 10 most relevant factors for classifying student performance.

Based on the assumptions made earlier, when looking for prevalence over the years, it can be said that the family's investment capacity has a much greater influence on student performance than color or race. However, this study does not deny that the poorest groups in Brazil are indigenous, black, yellow and brown, in that order.

When we rank the factors to identify the most prevalent inequalities, we find that indigenous people are the group most impacted by low income. This group is a minority in Enem and represents only 0.48% of registrations, according to Table 2. The second most affected group are blacks, who account for 10.91% of registrations.

With the exception of indigenous people, all comparisons of color and race are inferior to income inequality, evaluating only the top versus the bottom, because it is 3.9 times more difficult for someone at the bottom to excel than it is for someone at the top.

The social inequalities observable in relation to the type of school attended at primary level are greater for all combinations of color or race, with the exception of indigenous people. In 2022, it was 4.4 times more difficult for a public secondary school student to be labeled as "Above" than those enrolled in the private network.

Based on the latest information on primary education, it can be seen that the social inequalities observable from this information are greater than those obtained at secondary level or on color or race.

Even when we know the financial group in the Pareto

system, bottom or top, and combine the factors with levels, the pattern observed for the type of primary school, secondary school and color or race exposed above and in the previous sections remains the same. In general, the inequality of the groups increases whenever the bottom or top of any factor collected is compared.

From the factors analyzed, the most unequal or distant groups are indigenous people at the bottom compared to white people at the top (13.1 times more difficult to be labeled "Above"), those who chose Spanish and are from the bottom compared to those who chose English and are from the top (8. 4 times more difficult), the bottom group of public elementary school students versus the top group of private school students (7.5 times more difficult), the bottom group of public high school students versus the top group of private school students (6.6 times more difficult) and, finally, the bottom group whose color or race is black compared to the top group of white students (6.4 times more difficult).

All the graphs and analyses were generated on the basis of the answers given in the questionnaires by those who took the Enem exam. The results are not entirely precise about the type of school at primary and secondary level, due to unlabeled information, and only those who completed the stage entirely in a public or private institution were considered. However, these results can be considered a strong indication of the quality of education and student performance over two and a half decades, based on millions of pieces of data.

Finally, it can be said that all the educational policies of racial quotas and the inclusion of the poorest in universities and institutes [12] have had little or no qualitative effect over the years and that, for the vast majority of Brazilians, it is not possible to observe any approximation between social classes in relation to student performance, that is, Brazil has been and continues to be an unequal country, as analyzed in more than 102 million data points.

ACKNOWLEDGEMENT

We would like to express our gratitude to all those who believed in the potential of this research, as well as to our colleagues at IFMT (Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso) - *Campus* Barra do Garças and the funding institutions, FAPEMAT and CNPQ, for providing the necessary conditions for research in Mato Grosso, Brazil.

REFERENCES

[1] P. d. S. N. Lima, A. P. L. Ambrósio, D. J. Ferreira, and J. D. Brancher, "Análise de dados do Enade e Enem: uma revisão sistemática da literatura," *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, vol. 24, no. 1, pp. 89–107, 2019.

[2] F. L. d. Silveira, M. C. B. Barbosa, and R. d. Silva, "Exame nacional do ensino médio (enem): uma análise crítica," p. 1101, 2015.

[3] C. H. A. A. Moris, F. Casellato, M. M. Nascimento, G. Agostini, and L. Massi, "Distinção e classe social no acesso ao ensino superior brasileiro," *Tempo Social*, vol. 34, pp. 69–91, 2022.

[4] J. Franco, F. Miranda, D. Stiegler, F. Dantas, J. Brancher, and T. Nogueira, "Usando mineração de dados para identificar fatores mais importantes do enem dos Últimos 22 anos," in *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. Porto Alegre, RS, Brasil: SBC, 2020, pp. 1112–1121. [Online]. Available: https://sol.sbc.org.br/index.php/sbie/article/view/12867

[5] R. Koch, *O princípio 80/20: os segredos para conseguir mais com menos nos negócios e na vida*. Gutemberg, 2015.

[6] A. Boland, R. Dickson, and G. Cherry, "Doing a systematic review: A student's guide," *Doing a Systematic Review*, pp. 1–304, 2017.

[7] G. Dong and H. Liu, *Feature engineering for machine learning and data analytics*. CRC press, 2018.

[8] D. C. Montgomery, *Design and analysis of experiments*. John wiley & sons, 2017.

[9] G. de Oliveira Ribeiro, H. Zednik, and A. O. Nunes, "Recorte temporal e comparativo do desempenho dos estudantes de escolas públicas e privadas no exame nacional do ensino médio (enem) no município de fortaleza: o que a anova revela," *Concilium*, vol. 22, no. 6, pp. 648–662, 2022.

[10] F. M. Garcia, R. S. M. Caldas, and G. C. Torres, "O enem como política de avaliação e as contradições ao processo de democratização educacional," *Perspectiva*, vol. 39, no. 3, pp. 1–21, 2021.

[11] Ipea, "Ipea - Atlas da Violencia v.2.7." [Online]. Available: https://www.ipea.gov.br/atlasviolencia/filtros-series

[12] Brasil, "Lei n. 12.711, de 29 de agosto de 2012." [Online]. Available: $https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12711.htm$