INFORMÁTICA NA EDUCAÇÃO
teoria & prática
Vol. 26 | Nº 1 | 2023
ISSN digital 1982-1654 ISSN impresso 1516-084X
Páginas 110-117

# Analysis of the importance of social/racial quotas through ENEM's microdata mining

*Análise da importância das cotas sociais/raciais por meio da mineração de microdados do ENEM*

**Vandeir Vioti dos Santos**
Universidade Presbiteriana Mackenzie
professor.vandeir@gmail.com

**Pollyana Notargiacomo**
Universidade Presbiteriana Mackenzie
pollyana.notargiacomo@mackenzie.br

**Abstract**
Educational data analysis can provide new metrics and provide relevant information for creating tools. These seek to develop students' specific skills with the aim of increasing performance, improving existing methodologies and, mainly, developing educational policies. Given this scenario, a study was carried out with the objective of measuring the impact of the economic and social situation of Brazilian students. The National High School Exam (ENEM), in Brazil, is used by the students as one of the routes to apply and enter the higher education system, and also as a way to obtain scholarships. The main objective of this article is to emphasize the importance of the presence of racial and social quotas, through the analysis of ENEM microdata with Artificial Intelligence tools, interrelating their impact to the laws no. 12.711/2012 and no. 12.990/2014, not yet renewed in the governmental sphere, responsible for determining the rules regarding quotas in Brazil.

**Palavras-chave:** Social quotas. Racial quotas. Microdata mining.

**Resumo**
A análise de dados educacionais pode propiciar novas métricas e trazer informações relevantes para a criação de ferramentas. Estas buscam desenvolver habilidades específicas dos alunos com a finalidade de aumento de performance, aprimoramento de metodologias existentes e, principalmente, desenvolvimento de políticas educacionais. Visto esse cenário, um estudo foi realizado com o objetivo de mensurar o impacto da situação econômica e social dos estudantes brasileiros. O Exame Nacional do Ensino Médio (ENEM), no Brasil, é utilizado pelos alunos como uma das vias de inscrição e ingresso no sistema de ensino superior, e também como forma de obtenção de bolsas de estudos. O objetivo principal deste artigo é enfatizar a importância da presença das cotas raciais e sociais, por meio da análise dos microdados do ENEM com ferramentas de Inteligência Artificial, inter-relacionando seu impacto com as leis nº. 12.711/2012 e nº. 12.990/2014, ainda não renovado na esfera governamental, responsável por determinar as regras referentes às cotas no Brasil.

**Palavras-chave:** Cotas sociais. Cotas raciais. Mineração de microdados.

UFRGS PPGIE
PORTO ALEGRE
RIO GRANDE DO SUL
BRASIL

Recebido em: 13 de maio de 2023
Aprovado em: 07 de junho de 2023

INFORMÁTICA NA EDUCAÇÃO: teoria & prática

## 1. Introduction

With the technological advances, teachers can explore new strategies to innovate during their classes, for instance, by using technological resources (Leite et al. 2016). Data mining can reduce the challenges faced by professors (Gomes et al. 2017), since mining offers information that might influence the decision making of managers. This is already a reality for retailers, banks and insurance companies, and it might possibly help alleviate the challenges faced on the current education scenario (Gomes et al. 2017).

The analysis of ENEM (National High School Exam) microdata (Teixeira 2018) empowers the enhancement of the High School curriculum and the self-assessment of the students participation. Therefore, the goal of comparing data from 2013 to 2019 can present itself as a tool to be used while creating future educational plans and policies. Microdata supplied by INEP (Anisio Teixeira National Institute of Educational Studies and Research) is available to the general public, but in order to obtain any useful information from this data, it is necessary to dominate specific knowledge.

The education challenges, and the results in national assessments as the ENEM, allow for the creation of a series of studies focused on the development of new tools and methodologies that supports the process of teaching and learning. Among the educational data supplied by the Brazilian Government are: School Census, ENADE (National Students Performance Assessment), ENEM and Provão Saeb (Basic Education Assessment System). This study aims on the analysis of the data from the ENEM during the time course of 2013 to 2019, through the application of data analysis techniques (data mining).

Before the main research was conducted, a preprocessing was done. This preprocessing defined the algorithm to be used in main research by its results, and the objective was to find a prediction model for ENEM scores that would determine the results that the students would have in the Math and Writing exams (accuracy at least 80% for grades equal to or greater than 500). With this purpose, two ranges of grades were determined: less than 500 and greater than or equal to 500, with the highest possible score being equal to 1000. In order to accomplish the prediction, ENEM data from 2013 to 2019 were used. All common entries of all the years were used, and the cleanup of the database consisted of deleting records that did not had a Math score.

The preprocessing was focused in predictions using the Multi-Layer Perceptron (MLP) algorithms, Deep Learning for Java (DL4J), JRIP, PART, J48, RandomForest and RandomTree, all available in Weka. Tests were conducted comparing the performance of the algorithms mentioned. After not obtaining the target results, a field selection algorithm (InfoGainAttributeEval + Ranker) was used, the intent was to identify the inputs that mathematically had reduce influence (weight) in obtaining the final result. After applying this algorithm and performing new tests (using all the algorithms mentioned above), the target results were still not achieved. The algorithms that presented the best performances were MLP and DL4J. Dozens of hidden layers and neurons were used, but target results were not obtained.

The results secured during the preprocessing stated that the majority of people within the same socioeconomic background (monthly income per capita lower than 0.8 minimum wage in Brazil - R$1.212,00, approximately US$242.00 exchange rate quoted on 04/28/2023), in general, would obtain scores lower than 500 in Mathematics or Writing exams. After hundreds of tests, it was possible to observe that the values of missing target predictions were occurring due to people with lower socioeconomic conditions scoring grades above 500. The pattern of having a lower social economic background and grades higher than 500 was opposite to what the predictions of the MLP were showing. On account of this result, the main goal of the research was defined and the purpose was to discover the relation between the socioeconomic situation of the students and the performance in Mathematics and Writing tests in order to analyze the importance of social and racial quotas.

A quantitative study on this subject is pertinent, because a review of law no. 12,711/2012 was planned for the current year (2022), in Brazil. The legislation seeks through the mechanism of racial and social quotas, to promote more equal access to higher education and civil service examinations for people who are socially vulnerable.

From these perspectives, the present paper is organized as follows: section II presents an analysis of the data obtained from ENEM focusing on using these results to improve students' school performance; section III details materials and methods used on the present research; section IV discusses results of the analysis; lastly, section V consolidates conclusions and further works.

## 2. ENEM data analysis as a tool to improve school performance

The supply of microdata on education by the Brazilian Government provides an opportunity for researchers and everyone else involved in the education field to develop and enhance tools for the improvement of the quality of education. The technical difficulties on the manipulation and processing of microdata, due to the volume of it, makes this microdata very little explored, despite their relevance.

The systematic review study that analyzed articles whose central topic was the processing of microdata from ENEM or ENADE (Lima, Ambrósio, Ferreira, and Brancher, 2019), showed that the objective of the studies refers to improving the quality of education and students' performance, being the descriptive statistics the most used technique in data analysis. The authors suggest carrying out further studies using data mining techniques.

The analysis conducted through data mining techniques of ENEM's microdata (School Census of the years 2016, 2017, 2018) (Garcia, Rios-Neto, and Miranda-Ribeiro, 2021) concluded that the performance results of students from public schools in Brazil have a

similar performance with a 4.9% variation for more or less, and that investments in infrastructure can assist to improve the students' performance. A study conducted with ENEM's microdata, from 2019 on the State of Rio Grande do Sul, Brazil (do Carmo, Heckler, and de Carvalho, 2020), concluded that the socioeconomic situation and the social inequality are directly related to the students' performance. The study has also determined that students whose parents show a higher educational level are more prone to have better results.

The use of mining and clustering techniques associated with descriptive statistics to the 2019 ENEM microdata in the State of Minas Gerais (da Silva, Moreno, Gonçalves, Soares and Júnior, 2020) identified that the socioeconomic condition directly influences the performance of students in the exam, that indicates, low-income students tend to achieve unsatisfactory grades in ENEM. The study has also concluded that the performance of students from private and federal schools was similar.

The use of spatial statistics is also one of the tools used by researchers for data analysis. A study carried out in all of the Brazilian States with microdata from ENEM and the 2018 School Census, in addition to IBGE (Brazilian Institute of Geography and Statistics) statistics from 2018 (Melo, Freitas, Francisco and Motokane, 2022), corroborates that there is a strong indication that the mother's level of education, the family's economic situation, school infrastructure, race and the type of school the student attended (private or public) are directly correlated to the students' performance.

A study that used the Regression Tree Method and a model with 53 predictors with the 2011 ENEM microdata analyzed the variables related to the prediction of the Math grade (Gomes, Fleith, Marinho-Araújo, and Rabelo, 2021). The study concluded that the unsatisfactory performance in Mathematics is related to variables related to the student, his family and the school, some examples are family income, the student having studied in a public school, among others.

Researchers from the Federal University of Rio de Janeiro (Stearns, Rangel, Rangel, de Faria, Oliveira, and Ramos, 2017) carried out a survey to identify whether there was a possibility of making a prediction of the math grade through the socioeconomic data available in the ENEM microdata (2014) using statistical and regression tree techniques. The study concludes that there is a possibility of prediction and the methods that presented the best results were: Gradient Boosting and AdaBoost algorithms.

## 3. Materials and Methods

The data analysis method used in the research was divided in 10 steps (Figure 1), and each of these were applied individually to the data for each year researched (2013 to 2019).

In the first stage, the data used in the research was acquired (provided by INEP). The databases used has: 2013 (4.72GB); 2014 (5.68GB); 2015 (4.91GB); 2016 (5.29GB); 2017 (3.77GB); 2018 (3.19GB); 2019 (3.08GB); totaling 30.64 GB of analyzed data. The data collection was performed through the microdata portal of INEP (http://inep.gov.br/microdados).

Figure 1. Ten steps used in the methodology


Source: Prepared by the author.

The data was processed and transformed into a document-oriented database (MongoDB). After the cleanup on the MongoDB database the .csv files were exported to be used in Weka software (Waikato Environment for Knowledge Analysis) which was then used for the training of neural networks (MLP chose in preprocess phase with better results among the others tests).

The Weka was developed by the University of Waikato in New Zealand with the goal of obtaining information from the raw data. This software features a collection of machine learning algorithms aimed at data mining, performing tasks related to preprocessing, clustering, regression, classification and data visualization. The software is open source (Java) and is freely available under the GNU (general public license agreement) (Kiranmai and Laxmi). In 2005, the team that developed the software was awarded by the ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) (Russell and Markov 2006). Weka allows its users to compare different machine learning techniques and has become a tool used by academia and companies to carry out data mining researches (Hall et al. 2009).

The second stage consisted of analyzing the entries provided by each data file of each one of the years. The first data exclusion took place after this verification. Only the entries that were common to all years were kept. Focusing on a more detailed analysis, age and income data entries were divided into smaller ranges to better identify the influence of a given field on the studied objective (Math and Writing scores). For the same reason, special needs fields were grouped, as they had little information to be used individually. These changes allowed for a more detailed visualization of the data, despite having generated a much greater number of fields than the initial one.

The third stage of the process consisted on the exclusion the fields of registration data and the candidate's location. The fields referring to the registration number, city of residence code, name of the city of residence, federation unit of residence code, nationality, city of birth code, federation unit of birth code and the unit of birth abbreviation were excluded. Also excluded were the scores on Natural Sciences, Humanities, and on Languages and Communication Codes. All records that presented a blank Math grade were excluded.

The fields not excluded during the research were: abbreviation of the state where the student resides, gender, race, age, if the student has some kind of special needs, if the student has graduated high school, if the father of the student has graduated high school, if the mother of the student has graduated high school, if the father of the student has concluded a higher education degree, if the mother of the student has concluded a higher education degree, family income, if the student owns a computer, if the student owns a cell phone, if the student have access to the Internet, the type of school the student goes to (public or private), if the student has completed High School in the traditional form, or if he graduated High School through EJA (young people and adults education), quantity of people residing in the same house (required only to generate the income per capita) and if the house has a bathroom. The EJA is a teaching methodology, created by the Brazilian Federal Government to assist young people (over 18 years old), adults and elderlies that for any reason were not able to complete their studies. This teaching methodology allows the student to conclude the Middle School and/or High School in less time than it would normally take, in order to use these diplomas to have more professional opportunities and be able to apply for Higher Education opportunities.

In the fourth stage, the cleanup of the fields which had little or no influence on the expected result, was carried out in each of the years included on the research. Input selection algorithms (InfoGainAttributeEval + Ranker) and classification algorithms (PART, JRIP and MLP), were used with this purpose. The input selection algorithms allowed the identification of the fields that had the lowest weights for the final result. The use of PART and JRIP allowed the confirmation that the fields do not appear in the generated rules for the algorithm to make the prediction. The fields that appear in the PART and JRIP rules with at least 0.1% of the records were not removed.

After the removal of a field, the tests were repeated to check if there were any changes in the results. During this test phase, the MLP classification algorithm was used. Prediction percentage, prediction quality and confusion matrix data were checked before and after the removal of any field. If any significant and negative change occurred, the field was put back in the dataset.

At the end of the described process, the fields removed were: number of people living in the household, marital status, score of the five competency grades in the Writing test and special needs (sum of all fields that indicate special needs). It is noteworthy that the number of people in the household was later used to generate the income per capita.

The fifth stage of the process consisted on the cleanup of the duplicated records. After the cleanup of the fields, the records that were repeated were excluded in each one of the years.

The sixth step had the goal of validating the choice of the range of values chosen for the output (lower than 500 and higher or equal to 500). At first the choice was made because these values represented half of the maximum score possible and for being considered high enough to get approved in different Undergraduate Programs (scores equals to or higher than 500). In addition to the didactic component involved in choosing these score ranges, tests were executed to determine, by using technical criteria what would be the ideal score ranges for the outputs, using MLP, PART and JRIP algorithm and the confusion matrix. Initially several score ranges were created, for instance: Ranges of 100 points generating 10 different score groups. After the tests the ranges were reduced to 5 groups of registered grades, afterwards 3 groups and later a new test sequence concluded that the two ranges of scores used during the research, would be the chosen ones. The methodology applied while choosing the range of scores, consisted in testing each year of the data researched, analyzing the data referring to each one of the score ranges through the MLP, PART and JRIP algorithms and their respective confusion matrix results. In order to do that 70% of the data of each year was used for training purposes and 30% for tests. The tests were performed with all the years of this study, obtaining results without any discrepancies. Tests were also performed incorporating data from all years and again the results didn't show any discrepancies. For exemplification purposes, the confusion matrix with three groups of grades and a random sample with approximately 1.8M records. The group A consists in scores lower than 500, the group B has scores higher or equal to 500 and lower than 800, and group C presented scores higher or equal to 800. The confusion matrix results can be observed in Figure 2.

In the figure 2 it's possible to observe that group C do not show accuracy when compared to group A and B, that is, the elements of these range of scores were expressively classified as group B (7,608) and group A (250) instead of group C (zero). For this reason, only groups A and B were used during this research. Again, it's important to highlight that these results were tested with different score ranges of all years. All the tests were performed individually on each year and also with sample of data with distinct sizes, composed by information of all years.

Figure 2. Example of confusion matrix generated after analysis using MLP (Multi-layer Perceptron) field.

```
=== Confusion Matrix ===

       a       b       c   <-- classified as
 1239981  111817       0 |       a = A
  200616  276111       0 |       b = B
     250    7608       0 |       c = C
```

Source: Prepared by the author.

The seventh step was the inclusion of the writing scores as other criteria to be analyzed, in order to obtain a broader view of the academic aspect, in other words, the performance in the Humanities and the Exact Sciences would be analyzed.

The eighth step was the addition of the inputs related to family income. When observing the microdata of ENEM it was noticed that adopting the family income criteria, not necessarily would indicate the social

economic situation of a student. Therefore, the information of family income was divided by the number of residents in the house, resulting in the income per capita data, in addition of the family income.

The ninth step constituted on the automation of part of the testing (due to the quantity of necessary tests). The API (API = Application Programming Interface of Weka) was used, what allows tests to be performed directly in command line. Programs were created in order to execute the assemble of all the databases and the tests, what allowed the distribution of the tests in different computers, accelerating the process. Intermediate databases were created, each one was generated by year.

A process loop was created to systematically pulled out fields in order of importance, from higher to lower, to verify which fields, or group of fields, would replace the fields excluded in result of running JRIP and PART algorithms rules. The classification algorithms InfoGainAttributeEval + Ranker Where are used to classify the fields in an initial order of importance. After the classification, the most relevant field was systematically pulled out of the database and the classification algorithms PART and JRIP were used to confirm the results before and after the excluded field. The analysis was carried out studding the result rules from PART and JRIP. After the data was analyzed, it was possible to verify the most significant result fields (among the ones that still remained), and the process (described previously), was repeated. Following the new classification, the field of higher importance was again excluded and all the process redone. The process loop was only concluded when the quantity of records found in the prediction were reduced below 70% of the total. At the end of the described process, it was possible to rank and identify the fields and group of fields that had equivalence and a direct relation. Tests with fields in alternate order were performed to certify that the order of the fields would not change the results. After hundreds of tests, the fields race and social economic condition were emphasized.

On the tenth step forward, the test started to be performed directly on the database, using NoSQLBooster, robomongo and MongoDB-Compass software tools. The MongoDB-Compass allows visual and quick aggregations for testing. The NoSQLBooster allows testing queries quickly on screen and the robomongo allows javascript programming files to automate the tests.

The purpose of the queries was to assess the relationships between fields directly in the database. After all the tests were performed, the direct relation between race and socioeconomic conditions, with the results of Math and Writing scores was confirmed.

## 4. Results

The analysis of the data obtained shows that, when comparing the socioeconomic situation of the students and the race, it is possible to observe that the social economic situation presents as one of the main characteristics related to the academic performance of the student.

The law no. 12.711/2012 projects 50% of the enrollment quota of Federal Universities and Federal Institutes of Education, Science and Technology, to be reserved to students who attended high school fully in public schools and in the EJA regular courses, and from these quota 50% percent is reserved to students with family income per capita lower than 1.5 monthly minimum wages. Within these criteria the institutions must separate quotas for self-declared black people (black and brown) and indigenous people, on the same proportion of the population of these groups residing in the Federal units where the institution is located. All the other vacancies are open for free competition. The law no. 12.711/2012 foresees that within 10 years, that is in 2022, a new review must be performed on its content (de Oliveira et al. 2020). The law no. 12.990/2014 institutes the racial quota for Civil Service Examinations, on which 20% of the available quota for all positions are reserved for self-declared black people (Bulhões and de Oliveira Arruda 2020).

The studies regarding the use of social and racial quotas, in Brazil, when the publications completed on the Psychology field are analyzed, it's possible to notice that they are limited on the most part to exposing the opinion of people about this subject (de Oliveira et al. 2020). The main arguments of people interviewed in Brazil, against socioeconomic quotas, are that these positions should be conquered by meritocracy and that improvements on the basic education should be performed instead of the use of the quotas system (de Oliveira et al. 2020). In relation to the use of racial quotas, that by the law no. 12.711, 2012 (quotas law) is considered a socioeconomic quota, the people interviewed are majorly against having those opportunities reserved for racial quotas, with the argument that there is a social disparity and not a racial one (de Oliveira et al. 2020). The race quotas were established with the goal of fighting the institutional racism that exists in the Brazilian society (Bulhões and de Oliveira Arruda 2020).

The quotas law was a mechanism created by the Federal government with the goal of targeting these people and including them in the educational system (Ferreira and Guimarães 2021). The authors have come to the conclusion that after the creation of the quotas law, there was a historic increase, on the number of black people in the University. This law combines criteria, namely family income, color/race and the type of school where the student has concluded high school, to promote a more egalitarian access to higher education. Previously to the creation of this law, there wasn't a standard on the quota system to be adopted by universities and social institutes (da Silva Trindade and de Oliveira Mileo 2021).

This study raises the discussion, through the analysis of ENEM data, about the use of socioeconomic quotas as a way to promote access to universities, through the ENEM, for people with vulnerable socioeconomic situation.

The Tables 1 and 2, shows the data collection referring to grades lower than 500 in Mathematics or Writing according to the application of three different filters. The first filter (F1) is comprised by the information related to the type of school being public, the family income being lower than 4 minimum wage salaries and the fact that the student doesn't have a computer or

doesn't have a cell phone or doesn't have a bathroom in their residence. The filter number two (F2) refers to the students who have a family income per capita lower than 0.8 monthly minimum wage in Brazil. The third filter is a filter that uses information about the race (RF), it was used data from students who self-declared their race as any race other than white.

Table 1. Filters applied on grades lower than 500 in Mathematics or Writing

| Filter | 2019 | 2018 | 2017 | 2016 |
|---|---|---|---|---|
| F1 | 77.15% | 74.58% | 75.40% | 75.72% |
| F2 | 80.72% | 78.77% | 80.13% | 80.40% |
| RF | 62.94% | 62.10% | 62.12% | 61.42% |

Source: Prepared by the author.

Table 2. Filters applied on grades lower than 500 in Mathematics or Writing

| Filter | 2015 | 2014 | 2013 |
|---|---|---|---|
| F1 | 73.81% | 76.52% | 74.38% |
| F2 | 76.98% | 78.85% | 80.16% |
| RF | 59.75% | 58.86% | 57.66% |

Source: Prepared by the author.

When analyzing the data presented on this tables the filter F2 (family income per capita, lower than 0.8 monthly minimum wage in Brazil.) presented the highest percentage indicators in all years studied (2013 to 2019), which means that from the people framed in this filter, on average 79.43% (considering all the years studied), had grades lower than 500 in Mathematics or Writing. In contrast, the filter referring to the race (RF) presented the lowest percentage index in all the years studied, with an annual average of 60.69%, that is, comparing the average of filters F2 and RF it's possible to observe that the filter referring to the race (RF) had an average index of 18.74% lower than the filter F2. The filter RF shows its relevance while stays above average in all years.

On a second analysis, shown on the Tables 3 and 4, where there was a merge of filters F1 and F2 (denominated socioeconomic filter - SEF). Considering the average of the whole period studied 86.49% of the students who fit in the socioeconomic filter presented a grade lower than 500 in Mathematics or Writing, in contrast when observing the filter RF, referring to the race, the average percentage falls to 60.69%. It is possible to observe that the filter RF presented on average an index 25.80% lower than the socioeconomic filter.

Table 3. socioeconomic filter - SEF and race applied on grades lower than 500 in Mathematics or Writing

| Filter | 2019 | 2018 | 2017 | 2016 |
|---|---|---|---|---|
| SEF | 87.40% | 85.74% | 86.76% | 87.35% |
| RF | 62.94% | 62.10% | 62.12% | 61.42% |

Source: Prepared by the author.

Table 4. socioeconomic filter - SEF and race applied on grades lower than 500 in Mathematics or Writing

| Filter | 2015 | 2014 | 2013 |
|---|---|---|---|
| SEF | 84.92% | 86.38% | 86.91% |
| RF | 59.75% | 58.86% | 57.66% |

Source: Prepared by the author.

Table 5. Percentage of students who belong in the socioeconomic filter, in relation to the total of students who took the ENEM

| Filter | 2019 | 2018 | 2017 | 2016 |
|---|---|---|---|---|
| SEF | 79.61% | 78.51% | 80.31% | 82.47% |

Source: Prepared by the author.

Table 6. Percentage of students who belong in the socioeconomic filter, in relation to the total of students who took the ENEM

| Filter | 2015 | 2014 | 2013 |
|---|---|---|---|
| SEF | 80.40% | 82.20% | 81.35% |

Source: Prepared by the author.

Table 7. Grades greater than or equal to 500 in Mathematics and Writing and NOT framed in SEF filter

| Filter | 2019 | 2018 | 2017 | 2016 |
|---|---|---|---|---|
| NOT SEF | 44.96% | 47.91% | 46.01% | 48.14% |

Source: Prepared by the author.

Table 8. Grades greater than or equal to 500 in Mathematics and Writing and NOT framed in SEF filter

| Filter | 2015 | 2014 | 2013 |
|---|---|---|---|
| NOT SEF | 54.32% | 51.81% | 47.38% |

Source: Prepared by the author.

The Tables 5 and 6, shows the percentage of students who belong in the socioeconomic filter, in relation to the total of students who took the ENEM, it's possible to observe on average, considering the whole period studied, 80.69% of the students who took the ENEM test, belonged to this filter.
The average of students who didn't belong to the socioeconomic filter and had a score equal to or greater than 500 in Mathematics and Writing was 19.31% (100% minus 80.69% from Tables 5 and 6) and this are responsible for an average of 48.65% of the grades (in Tables 7 and 8).

Table 9. Filters applied on grades greater than or equal to 500 in Mathematics AND Writing

| Filter | 2019 | 2018 | 2017 | 2016 |
|---|---|---|---|---|
| F1 | 38.85% | 33.86% | 35.01% | 32.41% |
| F2 | 47.65% | 45.86% | 47.93% | 45.44% |
| SEF | 55.04% | 52.09% | 53.99% | 51.86% |
| FR | 40.00% | 41.06% | 41.18% | 39.58% |

Source: Prepared by the author.

Table 10. Filters applied on grades greater than or equal to 500 in Mathematics AND Writing

| Filter | 2015 | 2014 | 2013 |
|---|---|---|---|
| F1 | 28.05% | 33.41% | 32.64% |
| F2 | 39.41% | 41.01% | 46.31% |
| SEF | 45.68% | 48.19% | 52.62% |
| FR | 37.37% | 35.93% | 37.16% |

Source: Prepared by the author.

The Tables 9 and 10 shows the filters applied (F1, F2, SEF and RF) on grades equal to or higher than 500 in Mathematics and Writing. An important characteristic to take into account is that Tables 9 and 10 shows Mathematics and Writing grades that are equal to or higher than 500, the student obtained this score in both subjects.

## 5. Conclusions and Further Works

The analysis of ENEM's educational data, through data mining techniques, allowed the verification of the importance of using quotas as a way to reduce social inequalities. Of the total number of students who took the ENEM, about 20% did not fit the requirement of any racial or social quota, but despite representing a fifth of the participants, they were responsible for representing almost half of the grades equal to or greater than 500 in the Mathematics and Writing test (average of 48.65%) and considering that 86.49% of the students who fit in the social quotas did not reach the grade of 500 in these two subjects, it is possible to confirm through the data, the difference in the quality of education provided to different socioeconomic levels of the population. The study also concluded that socioeconomic quotas bring greater representation of educational inequality than racial quotas, but it is emphasized that historical issues involved are not in discussion. All the analyzes carried out in this research prove the importance of Brazilian laws, as the law no. 12,711/2012 and no. 12,990/2014 that seek to reduce social and racial inequalities, highlighting that in case of any changes during the process of reviewing the laws, they should be made in order to increase the promotion of equality and equity between people.

As a suggestion for future works, these article leaves two points to be explored. The first would be to carry out a study to define through data analysis what would be the ideal percentages to be applied in social/racial quotas. The second point to be researched is what other actions could be taken to mitigate the differences in the students learning level, such as a study on the adoption of an intelligent tutor system, consisting on a web tool that can be used in parallel to face-to-face teaching and that allows the student to learn interactive, flexible and personalized (Piramuthu 2005), as a way of helping to reduce this inequality.

## References

ARAUJO, J. N. DE F. L. et al. Um Catálogo de Recursos Educacionais Digitais (RED) Gratuitos de Matemática para auxiliar os professores do Ensino Fundamental. Anais do XXII Workshop de Informática na Escola (WIE 2016). Anais... In: V CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (CBIE 2016). 2016.

BULHÕES, L. M. G.; ARRUDA, D. DE O. Cotas Raciais em Concursos Públicos e a Perspectiva do Racismo Institucional. NAU Social, v. 11, n. 20, 2020.

DO CARMO, R. V.; HECKLER, W. F.; DE CARVALHO, J. V. Uma Análise do Desempenho dos Estudantes do Rio Grande do Sul no ENEM 2019. RENOTE, v. 18, n. 2, p. 378–387, 2021.

FERREIRA, I. D.; GUIMARÃES, C. H. S. A Efetividade das Cotas Raciais no Ensino Superior Público no Brasil Frente a Lei n 12.711/2012. Revista do Curso de Direito do Centro Universitário de Barra Mansa/UBM, v. 6, n. 1, p. 95–110, 2021.

GARCIA, R. A.; RIOS-NETO, E. L. G.; MIRANDA-RIBEIRO, A. DE. Efeitos rendimento escolar, infraestrutura e prática docente na qualidade do ensino médio no Brasil. Revista Brasileira de Estudos de População, v. 38, p. 1–32, 2021.

GOMES, C. M. et al. Predictors of Students' Mathematics Achievement in Secondary Education. Psicologia: Teoria e Pesquisa, v. 36, n. 38, 2020.

GOMES, T. C. S.; GOUVEIA, R. M. M.; BATISTA, C. M. Dados Educacionais Abertos: associações em dados dos

inscritos do Exame Nacional do Ensino Médio. Anais do XXIII Workshop de Informática na Escola (WIE 2017). Anais... In: VI CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (CBIE 2017). 2017.

HALL, M. et al. The WEKA Data Mining Software: An Update. SIGKDD Explorations, v. 11, n. 1, 2009.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Microdados ENEM. Disponível em: <http://portal.inep.gov.br/microdados>. Acesso em: 2018.

KIRANMAI, S. A.; LAXMI, A. J. Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy. Protection and Control of Modern Power Systems, v. 3, n. 1, 2018.

LIMA, P. DA S. N. et al. Análise de dados do Enade e Enem: uma revisão sistemática da literatura. Avaliação: Revista da Avaliação da Educação Superior (Campinas), v. 24, n. 1, p. 89–107, 2019.

MARKOV, Z.; RUSSELL, I. An Introduction to the WEKA Data Mining System. ACM SIGCSE Bulletin, v. 38, n. 3, p. 367–368, 2006.

MELO, R. O. et al. Impacto das variáveis socioeconômicas no desempenho do Enem: uma análise espacial e sociológica. Revista de Administração Pública, v. 55, n. 6, p. 1271–1294, 2021.

OLIVEIRA, I. A. DE ; VIANA, L. V.; LIMA, T. B. Cotas Raciais na Universidade: Uma Revisão Integrativa da Psicologia Brasileira. Revista Subjetividades, v. 20, n. Especial 1, 2020.

PIRAMUTHU, S. Knowledge-Based Web-Enabled Agents and Intelligent Tutoring Systems. IEEE Transactions on Education, v. 48, n. 4, p. 750–756, 2005.

SILVA, V. A. A. DA et al. Identificação de Desigualdades Sociais a partir do Desempenho dos Alunos do Ensino Médio no ENEM 2019 Utilizando Mineração de Dados. Anais do XXXI Simpósio Brasileiro de Informática na Educação (SBIE 2020), p. 72–81, 2020.

STEARNS, B. et al. Scholar Performance Prediction Using Boosted Regression Trees Techniques. ESANN, 2017.

TRINDADE, J. DA S.; MILÉO, I. DO S. DE O. Cotas Raciais para Negros no Ensino Superior brasileiro: Análise do Processo de Decisão. Revista Cocar, v. 15, n. 31, p. 1–21, 2021.