Analysis of ENEM's attendants between 2012 and 2017 using a clustering approach

Afonso Matheus Sousa Lima, Alexander Ylnner Choquenaira Florez, Alexis Iván Aspauza Lescano, João Victor de Oliveira Novaes, Natalia de Fatima Martins, Caetano Traina Junior, Elaine Parros Machado de Sousa, José Fernando Rodrigues Junior, Robson Leonardo Ferreira Cordeiro

Database and Images Group (GBDI)
Institute of Mathematics and Computer Science (ICMC)
Department of Computer Science – ICMC – University of São Paulo, Brazil
{afonso.matheus, alexanderchf, alexislescano, novaes.jvo, martins.nataliaf}@usp.br
{caetano, parros, junio, robson}@icmc.usp.br

Abstract. Data analysis is increasingly being used as an unbiased and accurate way to evaluate many aspects of society and their evolution over the years. This article presents an analysis of student's characteristics, between 2012 and 2017, in the most important exam for entry into higher education in Brazil, the *Exame Nacional do Ensino Médio* (ENEM). The intention is to gain insights of Brazilian regions, ENEM's areas of knowledge, type of school and accessibility, using a clustering method (K-means). An extensive and careful cleaning of the database was made in order to homogenize it and avoid types of statistical bias. The results of this work are presented objectively in the article, so it may be useful and used as a numerical base in works of socio-educational disciplines or studies that are interested in better understanding the evolution of ENEM in recent years. Finally, some discussions and restrictions on grouping results were presented in a timely manner.

Categories and Subject Descriptors: H.2 [**Database Management**]: Database Applications; I.5 [**Pattern Recognition**]: Clustering

Keywords: ENEM, KDD, Clustering, k-means, Elbow Method

1. INTRODUCTION

The *Exame Nacional do Ensino Médio* (ENEM) was created in 1998 and aims to evaluate the performance of Brazilian students who are finishing or have finished high school in previous years. This exam is also the main evaluation instrument for admission on higher education in Brazil, with more than 500 public and private universities that use it. It's so important in Brazil that it makes high school institutions, both public and private, shape their learning methodology by focusing on getting the best possible result in ENEM. Further, the exam is used as a prerequisite for the *Fundo de Financiamento Estudantil* (Fies), which finances higher education courses on private institutions and in the *Programa Universidade para Todos* (ProUni), for the concession of complete and partial scholarships in private higher education institutions. Nowadays, ENEM has been not only widely adopted all over Brazil but, since 2014, it's also used for the admission in institutions from other countries, such as the University of Coimbra and Algarve, in Portugal.

ENEM's format evolved since it first debut in 1998. The format considered in our work has its first edition in 2009, although small modifications were made year by year. Until 2017, ENEM was applied

This research was supported by FAPESP (São Paulo State Research Foundation) [grants 2018/05714-5, 2016/17078-0 and 2020/07200-9], CAPES (Brazilian Coordination for Improvement of Higher Level Personnel) [grant 001], and CNPq (Brazilian National Council for Supporting Research)

on two consecutive days, usually in November. ENEM assesses students in four areas of knowledge spread across these two days, with Social Sciences and their technologies and Nature Sciences and their technologies, on a Saturday and Languages, Codes and their technologies and Mathematics and their technologies on a Sunday. Each one of these parts consists of 45 questions of multiple choice, making it 90 questions per day, with a total of 180 questions. In addition, the exam includes an essay, which should be done along with another area of knowledge on the second day. In it, the candidate must write an thirty line essay, considering the proposed theme that changes with each new entry of the exam. Given ENEM's importance, over the years, the government has also focused on making the exam more accessible to an audience with some limitation or impediment, such as people with disabilities, for example. Currently the exam has inclusive policies with fourteen resources, such as braille exam, for visually impaired people, and Brazilian sign language translator-interpreter (*Libras*), for people with hearing impairment. Furthermore, special treatment is offered for breastfeeding women, and there exists an alternative version of the exam, with one similar level of complexity, for teenagers in the penitentiary system.

ENEM is directly related to the situation of higher education institutions in Brazil. Giving an overview of the situation of these institutions, data from last census of higher education[1], released in September of 2019, accounts 2537 higher education institutions. Of these, 88.2% (2238) and 11.8% (299) belong to the private and public network, respectively, with the private network being responsible for 93.8% of the total number of graduates vacancies in 2018, while the public network had a participation of only 6.2% in the total number of vacancies offered. Despite the last discrepancy in the distribution of vacancies, the best Brazilian universities are mostly from the public network and with that, are the most sought after by those taking the exam. This happens because in these universities students do not pay any fees and have several benefits to motivate their permanence, such as food subsidies and residence. As one can imagine, the small amount of public vacancies, added to the excellent performance of these universities, makes that only the best prepared students are able to be admitted.

Being an exam already consolidated, carried out for so long and taken by several students from all over Brazil, is essential that a comprehensive and well-founded analysis of its characteristics is made. Fortunately, the *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira* (INEP), a federal authority linked to the *Ministério da Educação* (MEC), provides data related to student characteristics and performances for each exam edition since 1998. These datasets include a wide variety of information, such as scores in each area of knowledge, type of school of the student, student residence location, if the student has any disabilities and many others. Given that and the lack of a comprehensive analysis in the literature, this article proposes an analysis of the students who have taken ENEM in a certain period of time, considering the time interval between 2012 and 2017. Besides that, was decided to separate students by geographic regions of Brazil, being then: South, Southeast, Midwest, Northeast and North. Each Brazilian region has significant cultural and economic differences between them. One may expect to obtain divergent knowledge from each of these regions. The approach of this article consists on using a clustering algorithm to group students based on their scores in each area of knowledge and in the essay, discovering similarly performing student groups for each Brazilian region. With this, we sought to answer the questions in the following:

(1) In this time, what is the performance of each student group in each region, related to their general average score in ENEM and in each area of knowledge separately?

(2) In this time, which type of high school institution, public or private, are predominant in each group?

(3) In this time, what are the characteristics of the students with disabilities?

[1] Available at: http://portal.inep.gov.br/censo-da-educacao-superior

This work is organized as follows: Section 2 discusses some related works that also investigated ENEM data; Section 3 presents the basic concepts necessary for the development of the analysis, in this case, these concepts are Clustering, K-Means and Elbow Method; Section 4 focuses on the investigation itself, presenting the databases that this work uses, the preprocessing that had to be done on those databases, the way in which the clustering was performed and the results obtained; Section 5 discusses the three main questions that were intended to be answered, as well as additional analyses that we intended to perform, but proved to be impractical and the reasons for that. Finally, Section 6 concludes the article with some remarks about this work and further research.

2. RELATED WORKS

Given the importance of ENEM in the context of Brazil, being the most important and predominant exam for the population to enter a high education institution, few studies approach different issues about it. These researches range from automatically solving the ENEM's questions [Silveira and Mauá 2018] to the use of ontology to allow better visualization of ENEM's data [Cabral et al. 2012]. Consequently, there are articles that use distinct techniques to gain non-obvious knowledge about the performance of Brazilian students in different entries of the exam [Viggiano and Mattos 2013; Simon and Cazella 2017; Leoni and Sampaio 2017].

In the article of Viggiano and Mattos [2013], the focus is on the performance of students in the 2010's exam considering each Brazilian geographic region. The authors use descriptive statistics as a method of analysis, focusing on the description of data, presented as graphs and tables. One of it's main contributions is the presence of three performance groupings of students in the year 2010: superior (South and Southeast), median (Midwest) and inferior (North and Northeast). Also, average scores of all regions indicate that Brazilian education has yet to go a long way to achieve acceptable minimum performance, principally in Nature's Sciences area.

Simon and Cazella [2017] generated a predictive model of the average performance indicator in Nature's Sciences area in the ENEM of 2015. Their method was based in a data mining approach, using a decision tree algorithm (j48) with the support of the Waikato Environment for Knowledge Analysis (WEKA)[2] tool to assist on each step of the knowledge discovery process. As results, the most important hierarchically independent variables for deciding the class of performance in Nature's Sciences area are the school type and the socioeconomic level. With that, the best performance groups are formed by students with medium to high socioeconomic level.

Leoni and Sampaio [2017] also used the 2015 ENEM's data to evaluate the performance of students from public and private schools in the southern region of Rio de Janeiro, using a clustering approach. The algorithm used was K-means, with k = 2, for the set of cities in the southern region of Rio. They sought to characterize the profile of schools and students in each group using the indicators of schools' size, municipality, administrative dependency and socioeconomic indicator. Overall, the similar performance standards between the two groups are sufficiently significant to support the existence of natural groupings among the types of schools for the region.

All these articles used data that have been made available by INEP, which provides information about students who have taken ENEM, hosting more than 10 years of exam entries. Nonetheless, there are limitations on previous related work that approach ENEM's data. They all analyze only one entry of the exam [Viggiano and Mattos 2013; Simon and Cazella 2017; Leoni and Sampaio 2017], and there is usually a database delimitation to focus on a specific region of Brazil or a specific area of knowledge [Simon and Cazella 2017; Leoni and Sampaio 2017].

An **evolutionary** analysis of student performance has yet to be done over a period covering a set of years in which the exams were offered. The Brazilian geographic division approached by Viggiano

[2] Available at: https://www.cs.waikato.ac.nz/ml/weka/

and Mattos [2013] is very interesting, given the cultural and economic differences of these regions, resulting in different statistics over the data. Both characteristics are also approached in our article.

3. BACKGROUND

This section presents some concepts necessary to understand the steps used in the methods applied in this work: Clustering, K-means and the Elbow Method.

3.1 Clustering

Clustering is a division of data into groups of similar objects [Berkhin 2006]. It's a very common task in data analysis, making splits in the database according to the similarity between the elements, that is, elements that are grouped together shall be similar between themselves and dissimilar in relation to the elements in other groups. Gan et al. [2007] exemplifies situations where clustering is well applied in real problems. One of them is in health care development systems, where cluster analysis is used to identify groups of people that may benefit from specific services. In this same context, another interesting example is in health promotion, where cluster analysis is used to select target groups that will most likely benefit from specific health promotion campaigns and to facilitate the development of promotional material for specific diseases. Clustering has also been used to explore information in other contexts, such as: data summarization, customer segmentation, social network analysis and even as a preprocessing step in many classification and outlier detection models [Aggarwal 2015].

An important part of the clustering process is the measure of similarity between database elements [Deza and Deza 2009; Paterlini et al. 2011]. Usually, the data used for clustering can be represented as a point in an n-dimensional space, e.g., geographic coordinates. The most common way to calculate the similarity between the elements is to calculate the distance between these points, where the distance becomes shorter as the elements are more similar. There are several distance functions for calculating the similarity between elements, each of them with it's specificity. Generally, they need to have some properties. Given a dataset $\mathbb{S}$ and a distance function $\delta : \mathbb{S} \; X \; \mathbb{S} \to \mathbb{R}^+$, is expected that δ respects the following properties:

(1) Identity: $\delta(s_i, s_i) = 0 \; \forall s_i \in \mathbb{S}$.
(2) Simetry: $\delta(s_i, s_j) = \delta(s_j, s_i) \; \forall (s_j, s_i) \in \mathbb{S}$.
(3) Not-Negativity: $0 < \delta(s_i, s_j) < \infty \; \forall (s_j \neq s_i) \; s_i$ and $s_j \in \mathbb{S}$.
(4) Triangle inequality: $\delta(s_i, s_k) \leq \delta(s_i, s_j) + \delta(s_j, s_k) \; \forall (s_i, s_j)$ and $s_k \in \mathbb{S}$.

Property 4 is not mandatory, but is very useful in some cases, especially in indexing processes. For cases that respect this property, the pair $< \mathbb{S}, \delta >$ is defined as a metric space [Deza and Deza 2009; Paterlini et al. 2011].

In addition to the distance functions, there are other similarity functions, such as: Bray-Curtis, Cosine and Jaccard. For these functions, when the higher the returned value, the more similar the input elements [Deza and Deza 2009]. The distance function used in our work is the Euclidean distance, which is one of the most used.

3.2 K-Means

K-means is one of the most famous clustering algorithms due to its simplicity and quality of results. It's used on several areas such as in data mining, pattern recognition, image analysis and others. The idea of k-means is to divide the n elements of the database into k groups, such that each group contains elements that have similarities between themselves, and dissimilarity with elements from other groups [Paterlini et al. 2011]. That is, k-means tries to generate groups so that the distance

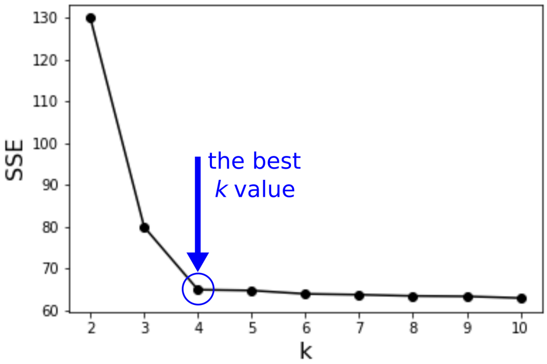


Fig. 1: Elbow Method

(similarity) between their elements and the corresponding group's center is minimized. Ortega et al. [2013] describes the main steps of k-means as follows:

(1) Select, randomly, k elements from the database to be the first centers.
(2) Calculate the distance from the database elements to the centers and group the elements with the nearest center.
(3) From the generated groups, re-calculate the centers of each group as the midpoint of the group.
(4) Steps 2 and 3 are repeated until some stopping criteria is reached, for example, the maximum number of iterations.

Typically, the algorithm expects as input the number k of groups and one dataset, where k has a large influence on the quality of the generated clusters. If k is too small, dissimilar elements may be on the same group and, if k is too large, elements with similar information may be on different groups, which implies groups with redundant information. In order to avoid this and other problems, several methods have been developed to define correctly the number of groups, one of the simplest and most well-accepted ones is the Elbow Method [Marutho et al. 2018; Bholowalia and Kumar 2014].

3.3 Elbow Method

the problem is to find the number of clusters that minimizes the sum of distances of each element to the center of its cluster. In this work, we used the Sum Square Error (SSE). SSE is the sum of Euclidean distance squared from each point to its cluster's center. If we select a large enough value to k, the function returns the smallest possible value. However, the best value for k is the one that generates the largest reduction in SSE (Equation 1). This statement is based on that while k increases, the SSE of the cluster starts to converge and from this point forward, increasing k does not reduce SSE value as much. Besides, bigger values of k increases the time execution, and groups with similar information will be generated. Then, the process of choosing the number of groups should be able to find the lowest k value that minimizes SSE. For this, the Elbow Method [Marutho et al. 2018; Bholowalia and Kumar 2014] follows a very simple strategy: it increments k until a large reduction of the SSE happens. Figure 1 exemplifies the process, which follows these steps:

(1) m sets of clusters are generated, such that $k = \{2, 3, \cdots, m\}$.
(2) for each set, SSE is calculated.
(3) a line graph is plotted, where the $x - axis$ is k and $y - axis$ is the corresponding SSE value.
(4) the point on the line which has the greatest slope or corner (from on, let us refer to it as elbow) is identified. This point is the ideal value of k.

$$\sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{1}$$

where C_i is i-th group and μ_i is the group center.

4. INVESTIGATION

This section details the steps to make the analysis proposed in our work. It constitutes of: Data Presentation, Preprocessing, Clustering and Results.

4.1 Data Presentation

INEP makes available, for consultation and download, databases generated by assessments, surveys and exams from their main educational programs[3]. In the ENEM section, during the period of this investigation, the databases are organized by year ranging from 1998 to 2018, each one has its own set of information about students who had taken the test in that year. For our work, only the databases from 2009 onward are of interest. This is justified because in 2009 ENEM was reformulated to unify the entrance exam of Brazilian federal universities[4]. Before that, ENEM's principle was to annually evaluate the learning of high school students across the country.

Unfortunately, some of the databases from 2009 onwards could not be used in this investigation. In some years, the format of the input file was very unorganized; as in 2011, where the data is arranged sequentially in a string for each record, with the identification of each value range present in an external document. There was also a lot of students that marked "*I don't want to inform*" on important information that we needed to answer one of the questions investigated in this article; as in 2018, where the information of the students' school type was mostly omitted. The structures of these databases can be updated by INEP at any time, so the situation described here refers to the date that these datasets were downloaded. Thus, we decided to exclude the years 2009, 2010, 2011 (as 2011 could not be used, to maintain a sequence of years, 2009 and 2010 were excluded too) and 2018 from our investigation.

With that being said, our analysis ranges from 2012 to 2017, a period that spans 6 years of exam editions. Metadata about these datasets can be seen in Table I. Although there are more than five million students in each dataset, some of them had to be removed so we could answer all the questions posed by this article, as explained later. Also note that there is a difference in the number of attributes in each dataset. The procedures adopted to filter irrelevant entries and structure these data are presented in the next section.

Year	**Before Preprocessing**		**After Preprocessing**	
	Num. of Students	**Num. of Attributes**	**Num. of Students**	**Num. of Attributes**
2012	5,791,065	80	1,205,203	36
2013	7,173,563	166	1,353,587	38
2014	8,722,248	166	1,449,074	36
2015	7,746,427	166	1,398,922	39
2016	8,627,367	166	1,489,705	39
2017	6,731,341	137	1,351,083	39

Table I: Information about datasets

[3] Available at: http://portal.inep.gov.br/web/guest/dados

[4] Available at: http://portal.mec.gov.br/busca-geral/179-estudantes-108009469/vestibulares-1723538374/13318-novo-enem

4.2 Preprocessing

The first step was to reduce the dimensionality of the datasets of each year. Initially, the data had 80 attributes for year 2012, 166 attributes for years 2013-2016, and 137 attributes for 2017. The most relevant attributes for the present research were carefully selected, i.e., those that are directly related to the research questions investigated, such as the grades, regions, attributes that determine a disability, among others. Attributes that were not of interest to this work were discarded, such as those that do not help to answer the research questions, or that had redundant values. It resulted in the projection of only 35 to 39 attributes for different years. To carry out this process, a line-to-line reading-and-preprocessing data approach was followed; this is because the large volume of data did not allow it to be loaded directly into main memory. The attributes used between all datasets can be seen in Appendix A.

To answer the proposed questions, we had to manually standardize the six databases. Is mandatory that certain attributes do not have empty values, mainly the attributes that describe aspects related to the performance of the students. All datasets have the attributes related to the score of the student in each area of knowledge of ENEM. These scores are the most important features for the clustering procedure, since student performance is measured by their results in the exam. Thus, we needed that all students in each database had scores for all areas of knowledge. So, we removed all students which had missing or 0 values in, at least, one area of knowledge. This procedure removed students that, either did not take the test on at least one day, or had their test canceled by some reason. Students who had no information about their school of origin were also removed; this is justified by the need to know if a student comes from a public or from a private school.

The result of this preprocessing step can be seen in the quantitative categorization presented in Figure 2. Figure 2(a) shows that, even after removing students from the original datasets, there is a clear pattern of participation by region along the years. For the time period considered, the Southeast is the predominant region in terms of participation. This can be justified by the fact that the Southeast is the Brazilian region with the largest number of inhabitants, encompassing the three most populous states in Brazil, i.e., São Paulo, Rio de Janeiro and Minas Gerais, according to the *Instituto Brasileiro de Geografia e Estatística* (IBGE)[5]. Second, is the Northeast region, followed by the South region, and ending with the Midwest and North regions with very similar statistics for all the years of this analysis. Figure 2(b) shows the participation of students separated by school type. For all the years, there is a very similar participation rate among the types of schools, always predominating the public schools over the private ones. This pattern was also expected, because most Brazilian schools are public, according to INEP statistics[6]. It's important to note that public schools in Brazil can be divided by their administrative dependency; municipal, state and federal dependencies exist. Lastly, Figure 2(c) shows the participation of students with special needs, such as visual impairment, motor disability, mental disability and others. As it can be seen, their participation tends to grow over the years.

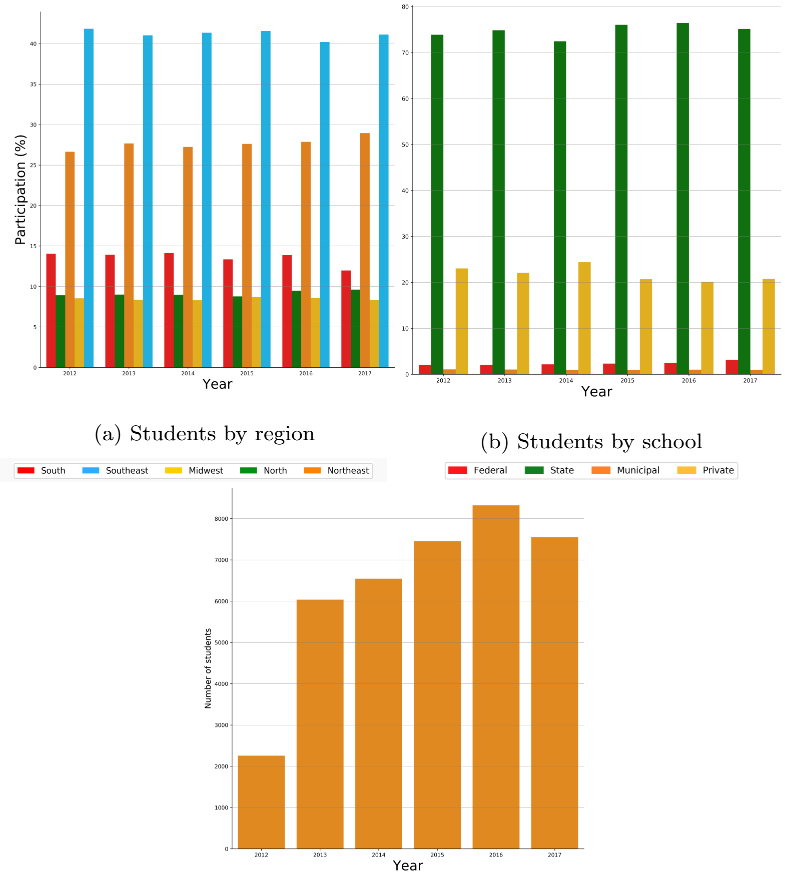
These data were then used by the clustering algorithm to define the performance groups, as described in the next section. With the selected attributes, it was possible to identify and report relevant characteristics about the students who are in each one of the groups.

4.3 Clustering

In order to gain non-obvious knowledge about students with similar characteristics in the time period considered, such as their general performance and participation as a whole, the datasets were divided into groups. Our intention is to make students with similar performance (considering both general

[5] Available at: `https://www.ibge.gov.br/apps/populacao/projecao/`

[6] Available at: `http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/dados-do-censo-escolar-mais-de-77-das-escolas-brasileiras-de-ensino-fundamental-anos-finais-sao-publicas/21206`



(a) Students by region

(b) Students by school

(c) Students with special need

Fig. 2: Information about datasets

performance and each area of knowledge) to be on the same group, providing the means for further analysis. The students were then grouped based on their performance on each ENEM's area of knowledge. Note that this is different than grouping students based on their general performance, which is the sum of each area of knowledge. For example, one student with 400 in Mathematics and 200 in Nature Sciences can have the same general performance than another student with 200 in Mathematics and 400 in Nature Sciences; but each one of them has difficulties on different areas of knowledge. Based on this fact, grouping students by the area of knowledge allows to analyze which area has the largest impact on the final results of the students.

For the task of clustering, we used the k-means algorithm. To select the ideal value of k, we used the Elbow Method. Both techniques were discussed in the previous Section 3. During the analysis, at the moment to apply the Elbow Method on the preprocessed ENEM's databases, depending on the year, the method returned different values for k, varying between 3 and 5. Because of that, though Elbow Method was used to help choosing the value of k, was decided to standardize $k = 3$ as the ideal value. This value was chosen because was a frequent value returned by the method in most years. This can be seen in Figure 3, representing the best k value (indicated by the vertical green line) for the year 2012, obtained with the SSE calculation, for each region. Furthermore, there is also a MEAN line in Figure 3, representing the mean SSE for this year, that is also equals three.

Another reason for a common value of$k$ is to facilitate comparisons over time. Therefore, for each year of this analysis, three groups per region were generated. To facilitate the discussion of the results, based on the general scores of the students in each group, the three groups were labeled as

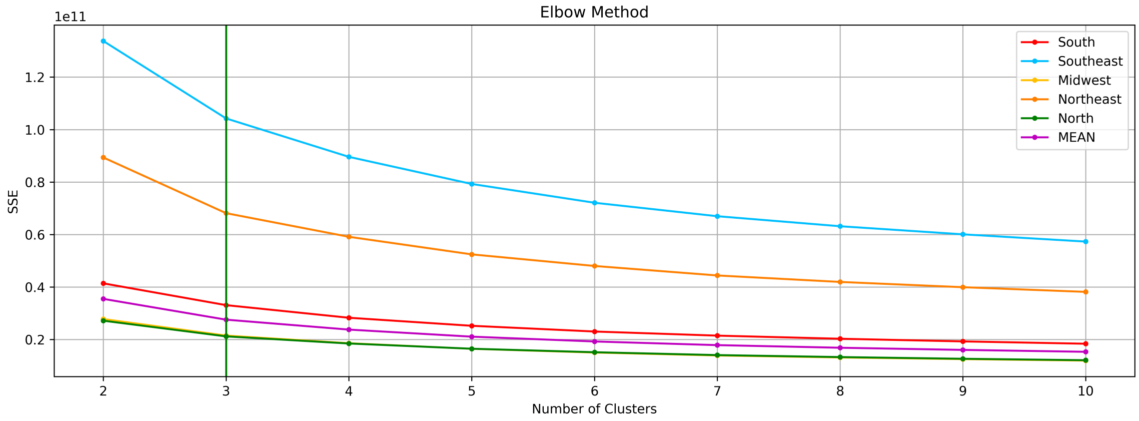
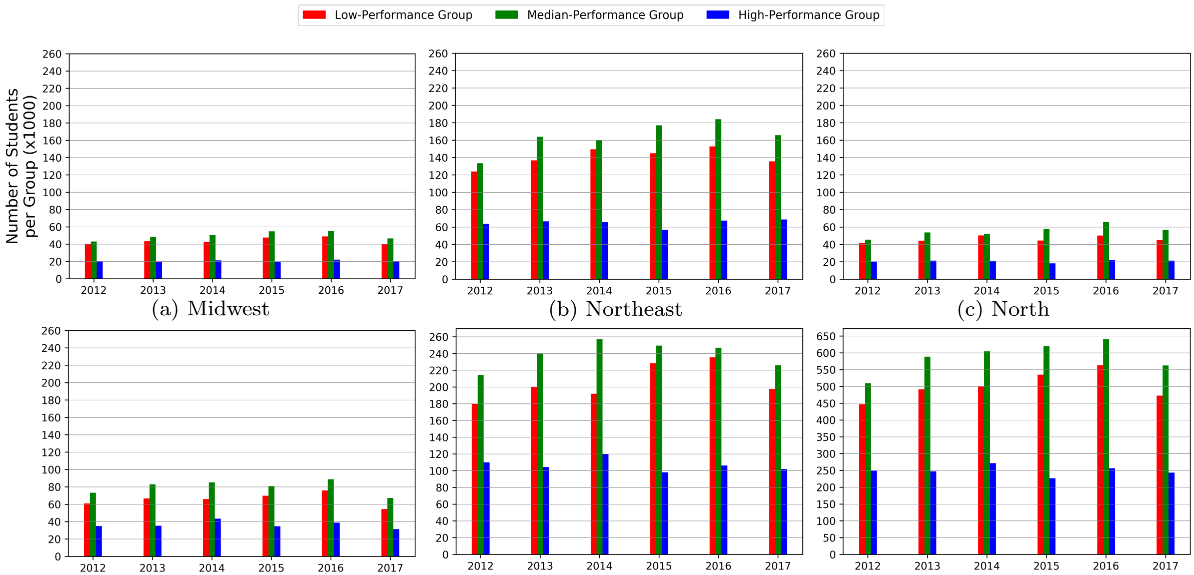


Fig. 3: Best k value for all regions in 2012



(a) Midwest (b) Northeast (c) North

(d) South (e) Southeast (f) All regions

Fig. 4: Evolution of the number of students in each performance group

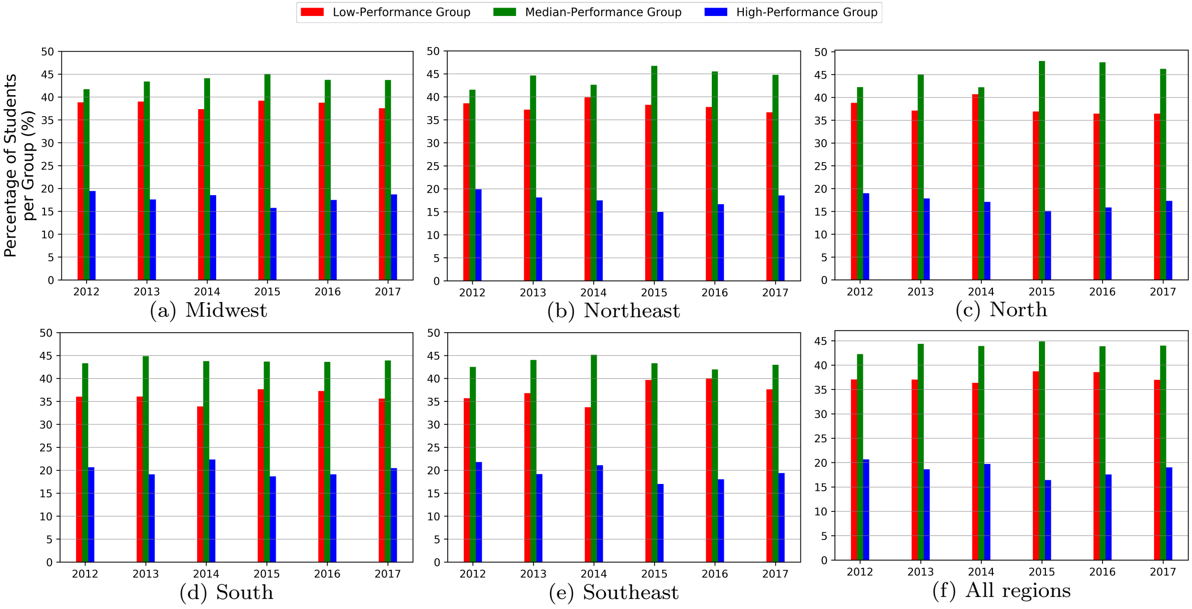
Low-Performance, Medium-Performance and High-Performance.

4.4 Results

To provide an overview of the clustering results, the number of students and the percentage of them in each performance group, separated by region, are showed in Figure 4 and Figure 5, respectively.

To answer the first question proposed by this article, which is: "*In the time period considered, what is the performance of each student group in each region, related to their general average score in ENEM and in each area of knowledge separately?*" (**Question 1**), a graph that shows the average score of the performance groups is plotted, in Figure 6. Besides that, the results on each area of knowledge are detailed in Table II.

To answer the second question: "*In the time period considered, which type of high school institution, public or private, are predominant in each group?*" (**Question 2**), the participation rate by school type is reported in Figure 7, still separating the students of each group by their region. Lastly, the third



(a) Midwest (b) Northeast (c) North

(d) South (e) Southeast (f) All regions

Fig. 5: Evolution of the percentage of students in each performance group

question: "*In the time period considered, what are the characteristics of the students with disabilities?*" (**Question 3**). The participation rate of people with disabilities by school type is showed in Figure 8. An average score graph for them was not plotted, since is known which performance group they are in, and with that information their average score were already presented (Figure 6). These results are discussed in the next section.

5. DISCUSSION

This section is structured in such a way that each one of the three proposed questions could be discussed. Further, other questions were considered by the authors of this work, but they were discarded due to missing information in the datasets used. These will also be discussed.

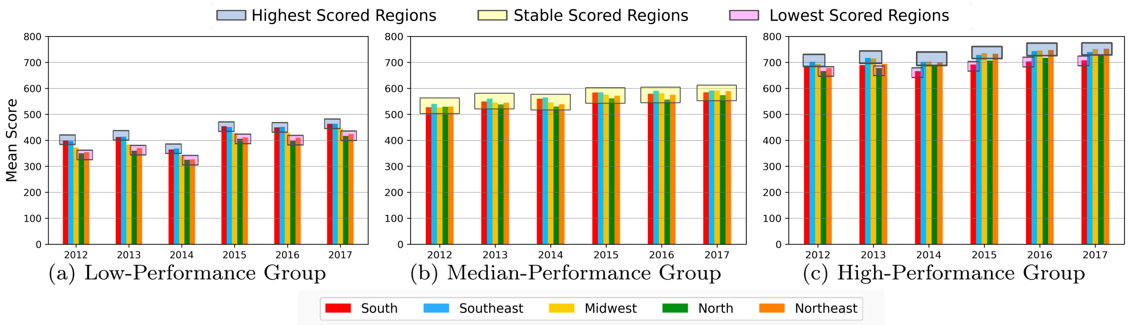
Before that, let us emphasize one general pattern that can be observed in Figure 4 and Figure 5: for all the time periods considered and for all regions, the majority of the students was grouped in the low and medium performance groups, while only a small number of students were grouped as high performers. This was already expected, as it confirms that the time period of our analysis, few students were able to perform very well in all competencies, which is a requirement to enter the more competitive graduation courses. It can also be seen that there was no relevant reduction in the gap between the number of students present in the low-median performance groups and that of the high performance group for the years of 2012 to 2017. As stated in Section 4.2, Southeast and Northeast are the most populated regions in the datasets used, which is also true when considering the real demographic density of these regions.

Area of Knowledge	**Low-Performance**		**Medium-Performance**		**High-Performance**	
	Min. Mean Result	**Max. Mean Result**	**Min. Mean Result**	**Max. Mean Result**	**Min. Mean Result**	**Max. Mean Result**
Nature Sciences	410,500	470,256	440,411	521,316	539,084	606,694
Social Sciences	442,157	514,409	492,930	582,901	595,283	652,275
Languages and Codes	413,223	484,662	467,144	551,292	548,176	609,109
Mathematics	395,344	459,418	430,212	551,177	562,025	685,635
Essay	325,404	463,576	526,298	591,400	665,876	752,426

Table II: Range of mean results by performance

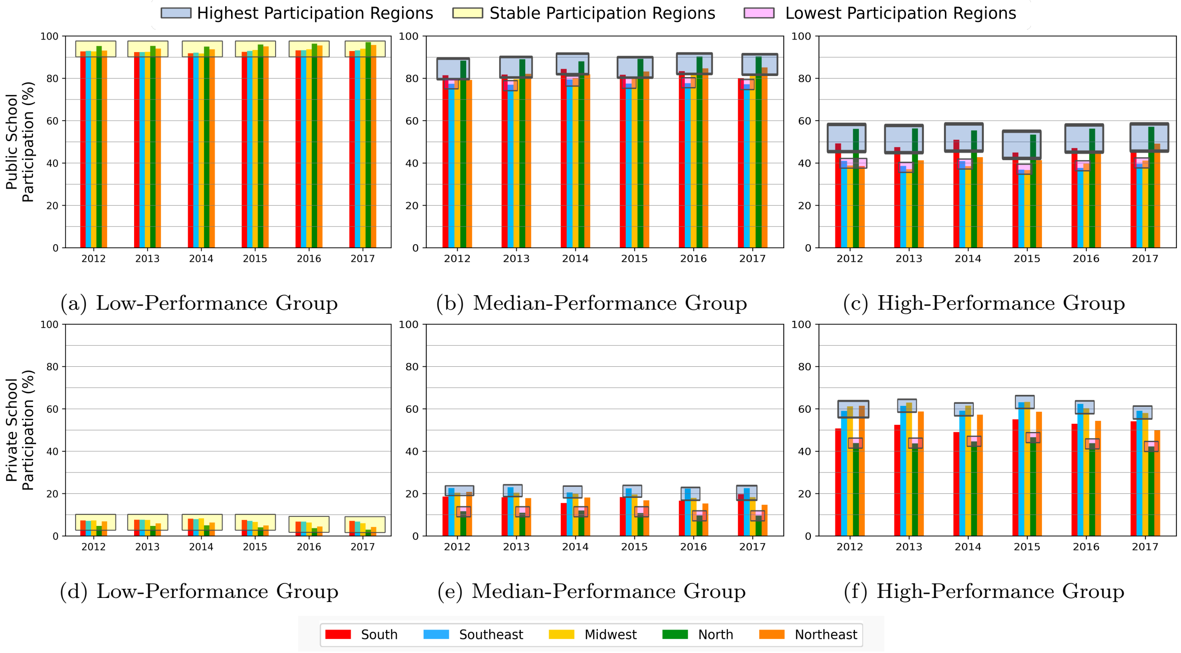
5.1 Question 1

Now, let us focus on the evolution of student performance between 2012 and 2017. In Figure 6, after 2014, the average score for all performance groups has increased, achieving the highest total average score in 2017. For this time period, the average grade goes from 400 to almost 500 points in the low-performance group, 500 to 600 points in the median-performance group, and 700 to 750 points in the high-performance group. Besides that, when considering the regions, some patterns can be recognized. In the low-performance group, South and Southeast regions have always the best performance, while North and Northeast have always the worst one. The median-performance group maintains a similar scenario in all regions for every year. However, the situation is a little bit different in the high performance groups. For the first years, Midwest and Southeast regions achieved the best scores; but in the later years, North and Northeast had caught on with them. Northeast improved every year and, in 2017, it has become one of the highest performance regions alongside with Midwest. Still on this matter, in Figure 7, the participation rate in the high-performance group of both regions, Northeast and Midwest, are one of the highest in public and private school types in 2017. This indicates that there is also a considerable number of high performers that come from these regions.



(a) Low-Performance Group (b) Median-Performance Group (c) High-Performance Group

Fig. 6: Total average group scores



(a) Low-Performance Group (b) Median-Performance Group (c) High-Performance Group

(d) Low-Performance Group (e) Median-Performance Group (f) High-Performance Group

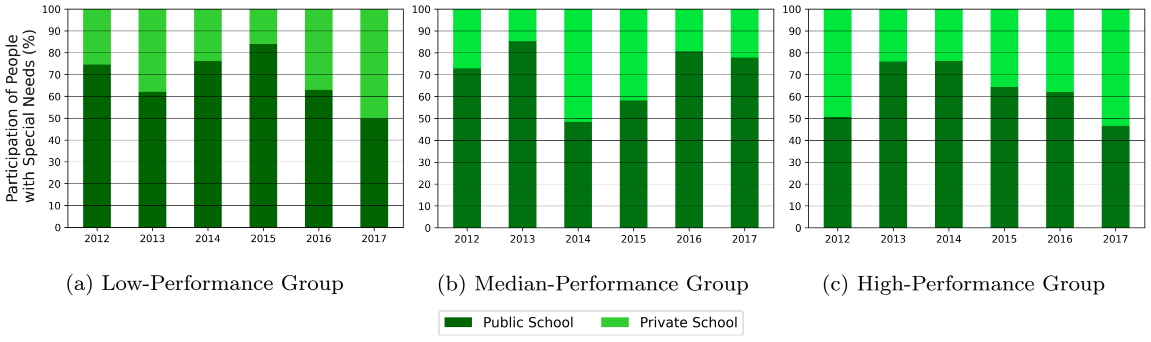
Fig. 7: Participation rate by school type

The performance of the students can be further analyzed by considering each area of knowledge independently. Table II reports the minimum and maximum mean scores for each area of knowledge in each performance group. The values for each year of the analysis are not shown because the difference between the years was not quite significant. The largest difference occurs in the minimum and maximum performance group mean scores for each area of knowledge, showing in which interval, over the course of six years, the students from each group stayed. With that in mind, Essay is the most discrepant among the performance groups. While the other areas of knowledge have relatively similar intervals in each performance group, with some of them even overlapping between these groups, the mean scores in Essay differs significantly. It's the worst minimal mean score in the low performance group, while it's the best one in the the high performance group. Therefore, having a high score in Essay can be a crucial determinant in the performance of ENEM as a whole.

5.2 Question 2

The participation rate by school type is one of the major problems in the Brazilian society. This reality is illustrated in Figure 7. Informed in Section 4.2, most schools in Brazil are public, and the low and medium performance groups are massively formed by public school students. In the low-performance group, only 10% of the students are from private schools. This scenario is repeated every year of the time period studied. On the other hand, for the high-performance group, there is a balanced participation rate between public and private schools, so is possible to do a more intrinsic analysis about it. Considering public schools, the North region always had the highest participation rate within the group. The South region placed second in this group in the first years studied, but the Northeast average score continually grew every year, going from last place to second in 2017. For private school students, the characteristics are quite different. Southeast and Midwest are the most predominant regions in participation along all the years studied.

Another pattern identified is the loss of participation of private school students in the Northeast region over the years, reaching the lowest point in 2017. For all years, the North region had the lowest participation in private schools. Thus, we may infer that public institutions in the South, Northeast and North are providing the best results for the majority of their students, with institutions in the Northeast focusing on improving their teaching methods year after year for ENEM. Private institutions with major participation rate are always located in Midwest and Southeast, indicating the presence of good private institutions that have maintained good results over the years. Unfortunately, for the low and median performance groups, students from public schools are by far the majority. This scenario applies to all regions of Brazil. At least, as discussed earlier, the average score of each performance group have increased over the years, which is very important for students in this group to be able to enter courses with not so elevated competition.



(a) Low-Performance Group (b) Median-Performance Group (c) High-Performance Group

Fig. 8: Participation rate of people with disabilities by type of school

5.3 Question 3

The participation rate of students with disabilities by the type of school is shown in Figure 8. Here, we shall discuss the results of the time period 2013-2017, since the year of 2012 has a very low number of impaired students (Figure 2). Most notably, there was a change in the predominance of the type of school in the high performance group. The participation rate in public schools in this group has been steadily decreasing, while the participation rate of the private ones has been increasing year after year. Such is that, in 2017, the private schools have surpassed the public ones in participation rate of students with disabilities. This could indicate that private high school institutions have increasingly invested in the education of these students, giving more support to them and applying teaching methodologies aimed at them. Furthermore, after 2015, the participation rate of public school students in the low performance group dropped, indicating that they migrated to one of the two other superior performance groups. A very small portion of the students in the dataset have the target profile of this analysis. This situation is somewhat alarming because, as in 2017, almost 24% of the Brazilian population have some kind of disability, according to UNESCO [7]. This may indicate that is still necessary to encourage the participation and to invest in the support of students with this profile, not only at the time of the exam, but throughout all their school life.

5.4 Impractical analyses

One of the questions initially proposed for this work was to calculate the number of students, from each year, who achieved results close to the value that would be achieved if the student had guessed all questions from some area of knowledge. Considering how the test is applied, each area of knowledge has 45 questions and each of them has 5 alternatives, being only one correct. With this, the probability of a student guessing a question correctly is 20% and, assuming a normal distribution, a representative guess grade interval could be found. However, ENEM uses a methodology applied by INEP that implements an anti-guess measure known as *Teoria de Resposta ao Item* (TRI)[8]. This methodolgy assigns weights to each question in the exam, based on the difficulty calculated by the hit rate of each one of them. For example, if a student gets the correct answer for a question, but many other students got it correctly too, this question is considered to be easy and has a low weight in the calculation of the score. Besides that, if a student gets a hard question correctly, but he/she misses an easy question, is understood that he/she guessed the hard question and it will be disregarded for the calculation of the score. Because of that, a deeper analysis of the questions needs to be done to make a representative guess grade interval, for which knowing the weight of each question is mandatory. But, unfortunately, the weights are not made available by INEP. Although a database scan can be done to try to infer these weights, it would be much more reliable if INEP itself updated the datasets informing the difficulty level of the questions, since other characteristics are taken into account by the TRI. This would be useful to make the grade calculation more transparent so that further analyses could be performed.

Another question that was considered for this work is to determine what the most common errors in the Essay are. That was initially considered because all datasets have an attribute that informs what were the students' errors in the Essay. Such errors could be: "*out of main topic*", "*orthographic error*", "*insufficient text*", and others. Unfortunately, for some of the datasets, the information about the errors on Essay does not seem to be correct. For example, in the year of 2012, all were given the same value, i.e., "*Present*", stating that the student was present at the time of the test, not reporting the actual mistake that the student made in the essay. Probably, this information may have been lost or even not collected properly, resulting in the setting of a default value for all students. Because of this, and due to our preference to analyze the longest possible time period, this question was discarded.

One last question that was considered refers to identifying the predominant social class in each of the

[7] Avaiable at: http://www.unesco.org/new/pt/brasilia/education/inclusive-education/persons-with-disabilities/
[8] Available at: http://portal.mec.gov.br/component/tags/tag/34530

performance groups. It would be quite interesting to know if, for example, the high performance group is predominantly made up of rich people. It would be a great indicator of Brazil's social inequality, as both private and public school students may have different social conditions. However, the only dataset that has an attribute with a social status rating is the one of year 2015. This is another information that INEP could provide to facilitate further analyses of the data. It's important to note that each dataset has a set of attributes related to questionnaire responses that may imply a student's social condition. These questions go from: "*Until what grade did your father or guardian study?*" to "*Does your house have a bathroom?*". Unfortunately, to infer a student's social class by using these questions, in addition to being quite complicated, is hampered by the fact that in almost every year the number of questions changes, as does the questions themselves.

6. CONCLUSION

This article proposed an analysis of the students who attended ENEM from 2012 to 2017 using a clustering algorithm, with data provided by a federal autarchy, known as INEP. Three groups of students were defined based on their performance, which was measured by the score in each area of knowledge of ENEM. These groups were identified as Low-Peformance, Median-Performance and High-Performance. For the time period studied, some notable changes were reported with the insight provided by the analysis. Some of them are: performance evolution of the groups, whith each Brazilian geographic region being analyzed separately; participation rate of students from public and private schools across the years, and evolution of participation rate of students with some disability. Besides that, some problems in the datasets were identified and reported using some other questions that could not be studied because of them.

As future work, further analyses could be performed in the context of ENEM. Most notably, ENEM underwent a redesign in 2018, with a shift in the order in which each area of knowledge is evaluated, as well as on days of test application, which are no longer over one weekend. Therefore, further analysis could be done to continue to investigate what was covered in this article, also verifying the impact of this change on the students' performance. One can also focus on inferring some of the questions that have been termed as impractical in our work, but this is not encouraged, since there is information missing that should be provided by INEP for the analysis to be valid and consistent.

REFERENCES

Aggarwal, C. C. *Data Mining: The Textbook.* Springer, Cham, 2015.

Berkhin, P. A Survey of Clustering Data Mining Techniques. In *Grouping multidimensional data.* Springer, Manhattan, New York City, USA, pp. 25–71, 2006.

Bholowalia, P. and Kumar, A. EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications* 105 (9): 17–24, 2014.

Cabral, S. P., Beduschi, N. B., Zancanaro, A., Todesco, J. L., and Gauthier, F. A. O. Aplicando Linked Data na Publicação de Dados do ENEM. In *ONTOBRAS/MOST.* Recife, Pernambuco, Brasil, pp. 176–181, 2012.

Deza, M. M. and Deza, E. *Encyclopedia of Distances.* Springer, 2009.

Gan, G., Ma, C., and Wu, J. *Data Clustering: Theory, Algorithms, and Applications.* Society for Industrial and Applied Mathematics, 2007.

Leoni, R. C. and Sampaio, N. A. d. S. Desempenho das escola públicas e privadas da região do vale do Paraíba: uma aplicação da técnica de agrupamentos Kmeans com base nas variáveis do ENEM 2015. *Cadernos do IME-Série Estatística* vol. 42, pp. 31–43, 2017.

Marutho, D., Handaka, S. H., Wijaya, E., and Muljono. The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. In *2018 International Seminar on Application for Technology of Information and Communication.* Semarang, Indonesia, pp. 533–538, 2018.

Ortega, J. P., Pires, C. E. S., Marinho, L. B., Mexicano, A., and Hidalgo, M. A. Early Classification: A New Heuristic to Improve the Classification Step of K-Means. *Journal of Information and Data Management* 4 (2): 94–103, 2013.

Paterlini, A. A., Nascimento, M. A., and Traina, C. J. Using Pivots to Speed-Up k-Medoids Clustering. *Journal of Information and Data Management* 2 (2): 221–236, 2011.

Silveira, I. C. and Mauá, D. D. Advances in Automatically Solving the ENEM. In *7th Brazilian Conference on Intelligent Systems (BRACIS)*. São Paulo, Brasil, pp. 43–48, 2018.

Simon, A. and Cazella, S. Mineração de Dados Educacionais nos Resultados do ENEM de 2015. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. Vol. 6. Recife, Pernambuco, Brasil, pp. 754–763, 2017.

Viggiano, E. and Mattos, C. O desempenho de estudantes no Enem 2010 em diferentes regiões brasileiras. *Revista Brasileira de Estudos Pedagógicos* 94 (237): 417–438, 2013.

APPENDIX A. DATA ATTRIBUTES USED

Attribute Name	**Description**
SG_UF_RESIDENCIA; NU_IDADE; TP_SEXO; TP_ESTADO_CIVIL; TP_COR_RACA	Attributes related to student information, such as: age, residence status, sex, among others.
TP_ESCOLA; SG_UF_ESC; TP_DEPENDENCIA_ADM_ESC; TP_LOCALIZACAO_ESC	Attributes related to school information, such as: type of school, location, among others.
IN_BAIXA_VISAO; IN_CEGUEIRA; IN_SURDEZ; IN_DEFICIENCIA_AUDITIVA; IN_SURDO_CEGUEIRA; IN_DEFICIENCIA_FISICA; IN_DEFICIENCIA_MENTAL; IN_DEFICIT_ATENCAO; IN_DISLEXIA; IN_DISCALCULIA; IN_AUTISMO; IN_VISAO_MONOCULAR; IN_OUTRA_DEF	Attributes related to the presence of a disability, such as: deafness, blindness, autism, dyslexia, attention deficit, among others.
TP_PRESENCA_CN; TP_PRESENCA_CH; TP_PRESENCA_LC; TP_PRESENCA_MT	Attributes related to the student's presence in each exam.
NU_NOTA_CN; NU_NOTA_CH; NU_NOTA_LC; NU_NOTA_MT; NU_NOTA_REDACAO	Attributes related to the score, between 0 - 1000, of each competence.
NU_NOTA_COMP1; NU_NOTA_COMP2; NU_NOTA_COMP3; NU_NOTA_COMP4; NU_NOTA_COMP5	Attributes related to the weighted score, between 0 - 200, of each competence.
TP_STATUS_REDACAO	Attribute that represents the situation of the essay, such as: no problems, blank, annealed, among others.