Socioeconomic analysis of students who took Enem between 2019 and 2022 using machine learning

Bruno S. Macêdo[1], Pablo S. Macêdo[2], Bruno H. Groenner Barbosa[3], Cristina M. Valadares[4], Patrícia M. Dias[4], Camila M. Saporetti[5], Leonardo Goliatt[6]

[1]*Systems and Automation Engineering Graduate Program, Federal University of Lavras*
Lavras, 37200-000, MG, Brazil
brunomsilva200@gmail.com
[2]*Chemistry Graduating, Federal University of Lavras*
pablocarola564@gmail.com
[3]*Dept. of Automatics, Federal University of Lavras*
brunohb@ufla.br
[4]*Dept. of Exact, State University of Minas Gerais*
Divinópolis, 37200-000, MG, Brazil
cristina.lima@uemg.br, patricia.dias@uemg.br
[5]*Dept. of Computational Modeling, Polytechnic Institute, State University of Rio de Janeiro*
Nova Friburgo, 28625-570, RJ, Brazil
camila.saporetti@iprj.uerj.br
[6]*Dept. of Computational and Applied Mechanics, Federal University of Juiz de Fora*
Juiz de Fora, 36036-900, MG, Brazil
leonardo.goliatt@ufjf.br

Abstract. The National Secondary Education Examination (Enem) is the exam that allows students, through the results obtained, to enter higher education institutions. Socioeconomic analysis is the means of evaluating the economic relationship with a portion of society. Through this analysis, which is carried out through socioeconomic questionnaires carried out in Enem, it is possible to analyze the factors that impact student performance. In this context, the objective of this work is to carry out a socioeconomic analysis of Enem from 2019 to 2022, aiming to identify possible social inequalities and factors that may influence students' performance in Enem. Due to the size of the Enem databases from 2019 to 2022, the following steps were adopted: selection, pre-processing, transformation, clustering and interpretation of the data through discovered knowledge. Furthermore, the development of the stages was supported by the Python programming language. As a result, between 2019 and 2022, two groups were divided for each year, a group of students with good performance and one of students with low performance.

Keywords: Enem, socioeconomic analysis, clustering, student performance

1 Introduction

The National Secondary Education Examination (Enem) is the exam that allows students, through the results obtained, to enter higher education institutions. There are three programs, called the Unified Selection System (Sisu), the University for All Program (Prouni) and the Higher Education Student Financing Fund (Fies), in which students can use their grades to obtain a scholarship or financing, in the case of private universities, and enter the case of public universities [1]. Through Sisu, students can enter public universities without any cost for their studies. Concerning Prouni, scholarships are granted for low-income students to enter private institutions. These scholarships can be partial or total, depending on the availability of the institution and the evaluation criteria adopted. Finally, Fies is a government financing program, in which students make payments after finishing their degree [1].

Students' performance on the exam allows education studies and indicators to be carried out in Brazil. Enem is of great importance to society, as most students seek to enter higher education through Enem. Due to the socioeconomic difficulties of some students, Enem is an exam that offers opportunities for students to enroll in undergraduate studies at a public or private educational institution, without having to pay the course fee or pay

only some part [2].

Socioeconomic analysis is the means of evaluating the relationship between the economy and society. Within this assessment, several factors are studied, such as financial issues, health, education, race, ethnicity, among others. It helps to understand the impact of these factors on population development and quality of life, as well as the social inequalities that exist in a given social group and a specific region, is important [3]. Even though Enem is not a means of evaluating the level of education in Brazil, through data obtained through socioeconomic questionnaires, when students register for an exam, it is possible to analyze the factors that impact student performance, and thus, outline strategies so that solutions can be sought for the development of education and students' entry into higher education [4].

It is clear, therefore, that analyzing Enem data can be a way of gaining insights, that is, an understanding of the impact of the socioeconomic situation on students' performance on the exam [5]. Due to the vast amount of information contained in the Enem database, carrying out analysis visually is unfeasible. To overcome this situation, automated computational techniques can be applied to assess differences between students efficiently and accurately. In this context, data mining has emerged as a tool for assisting in data analysis.

The socioeconomic analysis and identification of social inequalities in Enem is a topic that has been studied by several researchers. Various works have been carried out and proposed using computational intelligence techniques to understand the problem and seek solutions. Stearns et al. [6] analyzed the prediction of student performance in Enem 2014 based on socioeconomic data, focusing on mathematics grades due to their variation. The AdaBoost and Gradient Boosting regression algorithms were used, which were evaluated by Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) and R-Squared (R^2). Gradient Boosting obtained the best results, with MAE of 65.90 and (R^2) of 0.35. In turn, da Silva et al. [4] analyzed the social inequalities of Enem 2019 students using clustering algorithms and association rules, dividing them into two clusters. In Cluster A, the majority of students were from state schools, had a family income of less than R$2000, were predominantly black, were mixed race, were yellow or indigenous, and were mothers without secondary education. In Cluster B, the students come from several schools, but more than half are from state schools, predominantly white, with mothers who have at least secondary education, and 38.54% have a family income of less than R$2000.

Banni et al. [7] carried out an experimental analysis of Enem 2018, using data mining to identify the causes of student performance. The results indicate that students' performance in Enem is influenced by color/race, the region in which they live, and their parents' education. Students who choose English as a foreign language have better results. Macedo and Saporetti [5] used machine learning techniques to analyze data from Enem 2019 and 2020, verifying possible social inequalities between students in these years and predicting students' performance on the exam. The K-Means algorithm was used as a clustering technique to verify social inequalities. With the application of K-Means, 2 groups were obtained in the cluster, one composed of students with lower financial conditions and the other composed of students with better financial conditions.

Given all the information presented, this work aims to achieve the following objectives using computational intelligence techniques: conduct a socioeconomic analysis based on Enem data from 2019 to 2022; identify social inequalities reflected in Enem data from 2019 to 2022; perform a comparative analysis of social inequalities across Enem data from 2019 to 2022; identify the characteristics of students who achieved a certain performance in the exam.

The remainder of the manuscript is structured as follows: Section 2 describes the study area and the methodology used. Section 3 describes the conducted experiments and the achieved results. In Section 4, conclusions are presented.

2 Materials and Methods

2.1 Characterization of Datasets

The databases used in this work were Enem 2019 to 2022, which contain 5095171, 5783109, 3389832 and 3476105 samples, respectively. They have 76 attributes, such as sex, age group, nationality, marital status, color/race, type of secondary education, whether they have internet and a computer at home, monthly family income, the institution in which the student completed or will complete high school, and whether they are public or private, among others, and are available on the Inep website [8].

2.2 Clustering Method

K-Means is a clustering technique that divides a dataset into K different clusters [9]. Initially, K centroids of the clusters are randomly assigned or chosen from certain samples of the dataset [9]. These centroids are the

initial estimates for the centers of the clusters. Then, the K-Means algorithm aims to reduce the distance (*e. g.* Euclidean distance) between each data point and the nearest cluster, by updating the clusters' centroids. One of the challenges with K-Means is that not all values of K produce suitable clusters. Therefore, the algorithm is executed multiple times with different values of K, selecting those that offer the best interpretation of the clusters or the best graphical visualization, or even using some validation criterion to determine the best number of clusters [9].

2.3 Evaluation Metrics

The Silhouette Coefficient is a validation metric often used to determine the optimal number of clusters in clustering algorithms [10]. The silhouette coefficient measures an item's similarity to members of its own cluster (cohesion) compared to members of other clusters (separation). The silhouette validation technique involves calculating the silhouette coefficient for each sample, the average of these values for each cluster, and the overall average of the silhouette coefficient for the entire dataset. A high silhouette value indicates that an object is well integrated into its own cluster and distinctly separated from neighboring clusters. The silhouette coefficient is determined from the average of the intracluster distances a and the average of the distances to the nearest cluster b for each sample i. The silhouette coefficient is calculated as follows:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{1}$$

where a(i) is the average dissimilarity of the i-th object to all other objects in the same cluster and b(i) is the average dissimilarity of the i-th object with all objects in the nearest cluster. A silhouette value close to 1 indicates that the sample is well grouped in its cluster.

The Calinski-Harabasz Index, also known as the Harabasz variance ratio, is a metric used to evaluate the quality of clusters [11]. It is calculated as the ratio between the sum of dispersion between clusters and the sum of dispersion within clusters:

$$CH = \frac{Tr(B_k)/(k-1)}{Tr(W_k)/(n-k)} \tag{2}$$

where $Tr(B_k)$ is the sum of the variances between the clusters, $Tr(W_k)$ is the sum of the variances within the clusters, k is the number of clusters and n is the total number of samples. When groups are well separated, the dispersion between them is greater than the dispersion within them, thus resulting in a larger index.

3 Results and Discussion

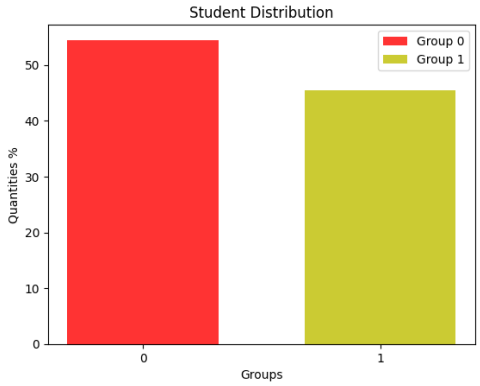
The computational experiments were carried out using the programming language Python, the libraries Pandas [12], Matplotlib [12] and Scikit-learn [12]. All experiments were performed on a computer with the following specifications: Intel(R) Core(TM) i5-8265U, 8 GB of RAM, and Windows 10 operating system.

To conduct the pre-processing analyses, all samples with empty values were removed, and only students who took the test on both days were considered. After the treatments, 909170, 520737, 592189 and 681900 samples were collected from the Enem 2019 to 2022 databases, respectively. To perform the group analyses, the K-Means clustering algorithm was used, with the number of groups varying from 2 to 5. The variation in the number of groups was based on the number of classes in the students' monthly family income, which was 5 classes (A, B, C, D, E). The number of groups adopted corresponds to the largest value of the Silhouette Coefficient and Calinski-Harabasz Index, for each variation in the number of groups. The students' grades on the exam, namely, the grades on CN (Natural Sciences and their Technologies), CH (Human Sciences and their Technologies), MT (Mathematics and its Technologies), LC (Languages, Codes and their Technologies), RE (Essay) and the average exam grade were normalized between 0 and 1, using the quantile_transform function from the Scikit-learn library, and the grades were the factors considered for grouping.

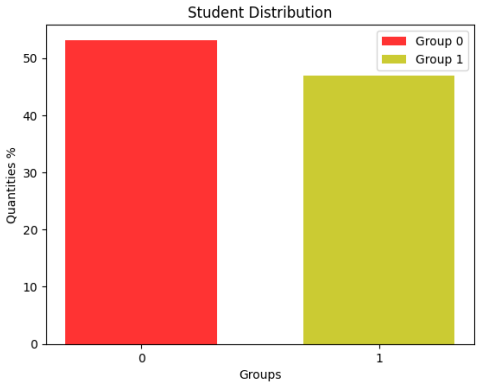
Two evaluation metrics were used to verify whether the ideal number of groups would be the same, the Silhouette Coefficient and the Calinski-Harabasz Index. The highest metric values obtained in the databases from 2019 to 2022 should represent the number of groups equal to 2. The graphics were constructed based on the number of groups with the highest metric value, which indicates a better representation of the clusters. The Silhouette

values were 0.464, 0.464, 0.458, and 0.441 for the years 2019 to 2022, respectively, while the Calinski-Harabasz Indices were 1186798.41, 682973.69, 749470.13 and 799014.22 for the same years, respectively.

Figures 1a, 1b, 2a and 2b show the distribution of students into two groups based on grades, from 2019 to 2022. Group 0 included students who did not perform as expected on the exam, while Group 1 included students who performed well on the exam.

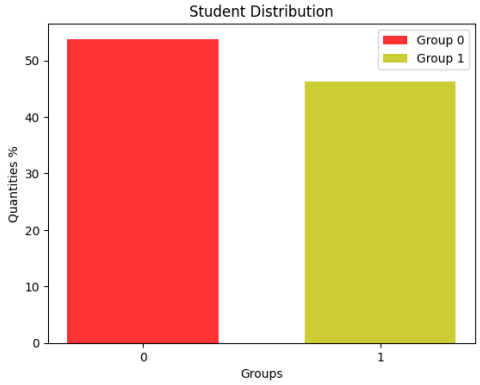


(a) Enem 2019.

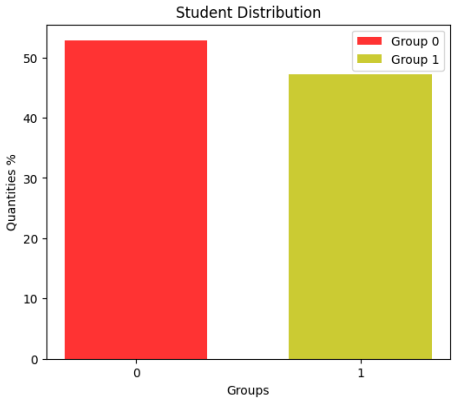


(b) Enem 2020.

Figure 1. Distribution of students in Enem 2019 and 2020.



(a) Enem 2021.



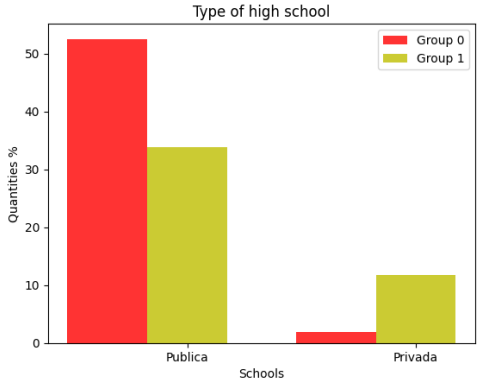
(b) Enem 2022.

Figure 2. Distribution of students in Enem 2021 and 2022.

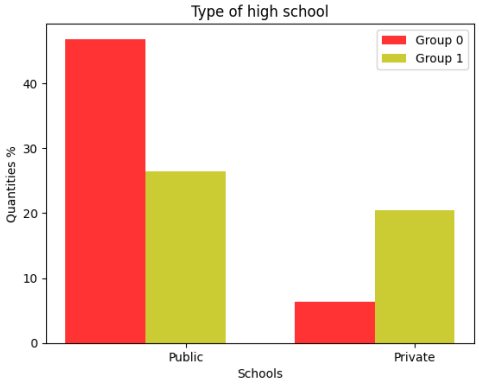
Figures 3a, 3b, 4a and 4b show the division of groups concerning the type of high school the students attended. According to the data from 2019 to 2022, the majority of students who did not perform as expected on the exam, represented by Group 0, were from public schools. The students in Group 1, which represents students with good performance, had some who studied in both public and private schools.

Figures 5a, 5b, 6a and 6b show the groups in relation to the type of color/race of the students. It can be seen that in 2019 to 2022, a large proportion of students with unexpected performance in the exam, represented by Group 0, were brown in color/race. The students with good performance, represented by Group 1, are mostly white. It is also clear that indigenous color/race has the least impact on the graphs, largely due to the number of indigenous people with access to education.

Finally, Figures 7a, 7b, 8a and 8b show the graphs in relation to monthly family income from Enem 2019 to 2022. In the years 2019 to 2022, Group 0, represented by students who did not perform as expected in the exam, had a maximum monthly family income of two times the minimum wage, that is, belonging to class E. In Group 1, represented by good performing students, most students were also represented by class E. However, for the years 2020, 2021 and 2022, there was an increase in good-performing students for classes C (4 to 10 minimum wages) and D (2 to 4 minimum wages). Furthermore, the difference between the number of students in classes C (4 to 10 minimum wages), D (2 to 4 minimum wages) and E decreases for Group 1.

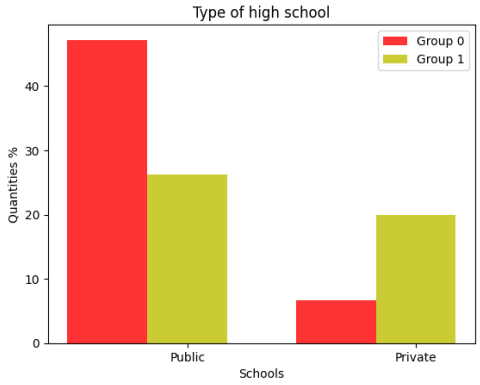


(a) Enem 2019.

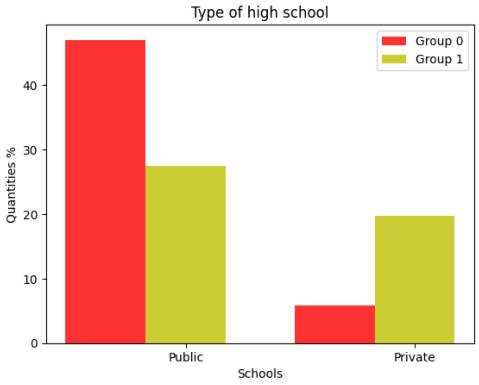


(b) Enem 2020.

Figure 3. Type of high school in Enem 2019 and 2020.

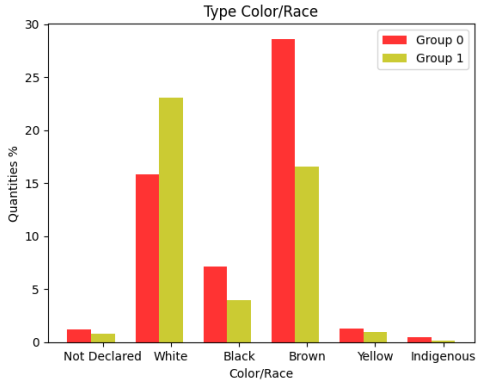


(a) Enem 2021.

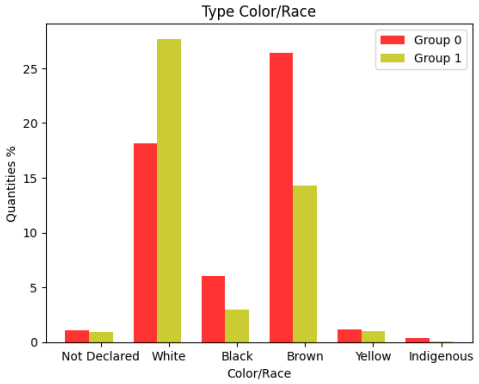


(b) Enem 2022.

Figure 4. Type of high school in Enem 2021 and 2022.

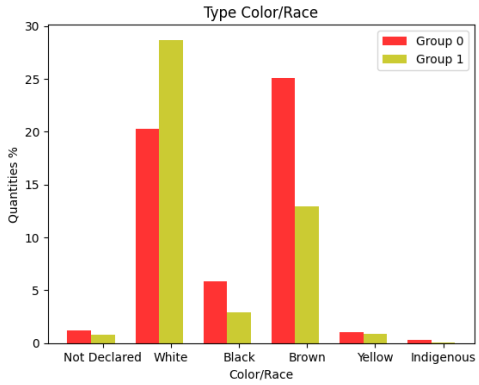


(a) Enem 2019.

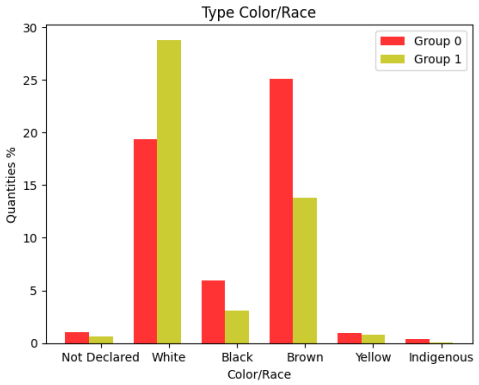


(b) Enem 2020.

Figure 5. Categories of color/race in Enem 2019 and 2020.



(a) Enem 2021.



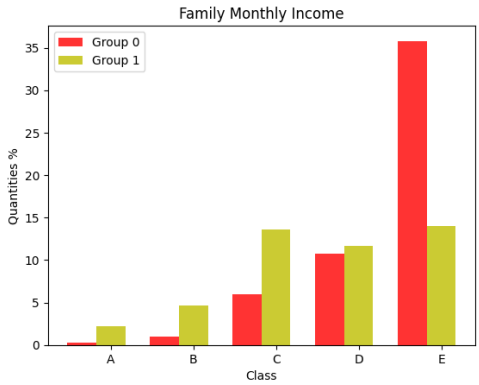
(b) Enem 2022.

Figure 6. Categories of color/race in Enem 2021 and 2022.

(a) Enem 2019.

(b) Enem 2020.

Figure 7. Categories of the class monthly family income in Enem 2019 and 2020.



(a) Enem 2021.

(b) Enem 2022.

Figure 8. Categories of the class monthly family income in Enem 2021 and 2022.

In relation to the analyses carried out, the majority of the students who performed well in Enem from 2019 to 2022, were from private schools, had a white color/race, had a monthly family income in class C in which the minimum wage was 4 to 10, class D in which the minimum wage was 2 to 4, and class E in which the minimum wage was up to 2. Even so, of the students who did not perform as expected, the vast majority were from public schools, had a brown color/race, and had a monthly family income of class E, which is up to 2 times the minimum wage.

4 Conclusions

In this work, it was identified that the student's type of school, color/race and the student's monthly family income impact their performance on the exam. Another point is that the difference between the number of students who signed up for the exam and those who took the Enem on both days was quite large. Various factors may have led to this decrease, as many may not be able to access the place where the exam is carried out, as some may have given up on taking the exam as well, among other factors. As future experiments, predictions of the grades obtained by students from Enem 2019 to 2022 will be conducted using Machine Learning models, such as Artificial Neural Networks (ANN), Random Forest (RF) and Support Vector Machine (SVM).

Acknowledgements. The authors would like to thank the State University of Minas Gerais (UEMG), the Federal University of Lavras (UFLA), the Postgraduate Program in Engineering Systems and Automation (PPGESISA), the Minas Gerais State Research Support Foundation (FAPEMIG) for the support given and the National Institute of Educational Studies and Research Anísio Teixeira (INEP) for providing data for this work.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] C. D. F. WESTPHALEN-RS and A. S. DE VARGAS. As políticas públicas para a educação superior no brasil pós ldb/96: O enem, sisu, prouni e fies e suas (des) continuidades, 2021.

[2] R. Travitzki, M. E. Ferrão, and A. P. Couto. Desigualdades educacionais e socioeconômicas na população brasileira pré-universitária: Uma visão a partir da análise de dados do enem. *Education Policy Analysis Archives*, vol. 24, pp. 74–74, 2016.

[3] dos A. P. G. Santos, T. F. G. Motti, da J. A. G. G. Silva, and M. A. S. Francelin. A importância do estudo socioeconômico para a equipe interdisciplinar em saúde auditiva. *Serviço Social & Realidade*, vol. 22, n. 2, 2013.

[4] da V. A. A. Silva, L. L. O. Moreno, L. B. Gonçalves, S. S. R. F. Soares, and R. R. S. Júnior. Identificação de desigualdades sociais a partir do desempenho dos alunos do ensino médio no enem 2019 utilizando mineração de dados. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pp. 72–81. SBC, 2020.

[5] B. d. S. Macedo and C. M. Saporetti. Analysis of the impact of the pandemic on social inequalities in enem 2019 and 2020 using machine learning. *Semina: Ciências Exatas e Tecnológicas*, vol. 44, pp. e48234, 2023.

[6] B. Stearns, F. Rangel, F. Firmino, F. Rangel, and J. Oliveira. Prevendo desempenho dos candidatos do enem através de dados socioeconômicos. In *Anais do XXXVI Concurso de Trabalhos de Iniciação Científica da SBC*. SBC, 2017.

[7] M. R. Banni, M. V. d. P. Oliveira, and F. C. Bernardini. Uma análise experimental usando mineração de dados educacionais sobre os dados do enem para identificação de causas do desempenho dos estudantes. In *Anais do II Workshop sobre as Implicações da Computação na Sociedade*, pp. 57–66. SBC, 2021.

[8] INEP. Microdados do enem. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em: `https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem`, 2024.

[9] K. P. Sinaga and M.-S. Yang. Unsupervised k-means clustering algorithm. *IEEE access*, vol. 8, pp. 80716–80727, 2020.

[10] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 1987.

[11] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, vol. 3, n. 1, pp. 1–27, 1974.

[12] D. Y. Chen. *Pandas for everyone: Python data analysis*. Addison-Wesley Professional, 2017.