CIENTIFIC Publishing Limited

Digital Technologies Research and Applications

http://ojs.ukscip.com/index.php/dtra

Article

Determinants of Performance in the Public Schools of the State of Rio de Janeiro, 2023

João Vitor Gomes Rodrigues [1], Rui Alexandre Castanho [2,3,4], Rodrigo Franklin Frogeri [2] and Fabricio Pelloso Piurcosky [2,*]

[1] Data Science and Analytics, University of São Paulo, São Paulo 13418900, Brazil

[2] Integrado Educational Center, Campo Mourão 87300410, Brazil

[3] VALORIZA –Research Centre for Endogenous Resource Valorization, Polytechnic Institute of Portalegre, 7300186 Portalegre, Portugal

[4] Advanced Research Centre, European University of Lefke, 99010 Mersin, Turkey

[*] Correspondence: fabricio.pelloso@grupointegrado.br

Received: 10 February 2026; **Revised:** 13 March 2026; **Accepted:** 27 March 2026; **Published:** 13 August 2026

Abstract: This study investigates how socioeconomic, demographic, and institutional factors are associated with the performance of 180 public school proxies in Rio de Janeiro on the 2023 National High School Exam (ENEM), utilizing official Anísio Teixeira National Institute for Educational Studies and Research (INEP) microdata. The analysis is based on the application of multiple linear regression models with robust HC3 standard errors and unsupervised clustering techniques. To mitigate multicollinearity issues between parental education and household income, a Socioeconomic Status Index (SSI) was constructed via Principal Component Analysis (PCA), explaining 92.19% of the joint variance of these variables. Regression results identified three consistent determinants: the SSI showed a direct positive association with scores; racial composition, where a higher proportion of non-white students correlated with lower performance; and administrative dependency, with federal schools significantly outperforming state schools. Complementarily, K-means cluster analysis revealed four distinct institutional profiles, showing a gap of approximately 70 points between performance extremes. Robustness tests, including log-linear regression and bootstrap validation with 1,000 resamples, confirmed the stability of the estimates. The findings indicate that the performance of public schools in Rio de Janeiro reflects not only structural socioeconomic inequalities but also a marked heterogeneity across educational networks and racial profiles. It is concluded that overcoming these bottlenecks requires focused, evidence-based educational policies capable of addressing the territorial and institutional disparities that persist in the state's public education system.

Keywords: Educational Inequality; Principal Component Analysis; Multiple Linear Regression; Clustering; ENEM Microdata

1. Introduction

In the state of Rio de Janeiro, public schools located in close geographic proximity may exhibit differences exceeding 70 points in the National High School Exam (ENEM). This dispersion, reflected in institutions' average performance, shapes individual life trajectories: it may determine access to a prestigious public university or, alternatively, exclusion from higher education. More than a contingent outcome, it constitutes an indicator of the fractures that permeate Brazilian society and are reproduced across successive generations. Identifying the determinants of this phenomenon is therefore a necessary condition for guiding strategies aimed at reducing educational asymmetries [1].

In this regard, the economics of education has consolidated, over recent decades, a theoretical framework that recognizes socioeconomic, demographic, and institutional factors as key determinants of academic performance [2]. In Brazil, the ENEM has become one of the principal instruments for addressing this issue, particularly following its reformulation in 2009 [3].

The theoretical foundations of these analyses can be traced back to seminal studies in the sociology of education. The report established a conceptual milestone by demonstrating that family characteristics—such as parental education, household stability, and social capital—exert a more substantial influence on academic achievement than the resources available within schools themselves. This finding redirected analytical attention toward socioeconomic conditions [4].

Human capital theory, provides a substantive complement by conceptualizing education as an investment in individual productivity. Within this framework, families with greater resource availability tend to allocate higher levels of investment to the development of their children's cognitive skills. Consequently, more advantaged socioeconomic contexts provide not only material resources but also stimulating environments that foster academic development [5].

Soares and Alves [6], in turn, introduced the concept of cultural capital, demonstrating that family-based dispositions—such as reading habits, command of formal language, and familiarity with school codes—generate systematic advantages or disadvantages within the educational process, transmitted in subtle and continuous ways.

In the Brazilian context, an extensive body of empirical literature corroborates this explanatory framework. Franco et al. [7], using data from the Basic Education Assessment System (SAEB), identified systematic associations between academic performance and variables such as parental education, household income, and the possession of cultural goods. Melo et al. [8] showed that the average socioeconomic status of schools influences educational outcomes even after controlling for students' individual characteristics, thereby evidencing composition effects that amplify initial inequalities. More recent investigations confirm the persistence of these patterns. Jollife and Cadima [9] demonstrate that socioeconomic factors explain a substantial share of the variation in ENEM scores.

These findings reinforce the need for analytical approaches capable of capturing the complexity of interactions among individual, family, and institutional factors. Recent methodological advances have enhanced the precision of such analyses. Principal Component Analysis (PCA) has become established as an appropriate technique for synthesizing multiple socioeconomic variables into parsimonious and statistically valid indices [10]. Simultaneously, cluster analysis methods enable the identification of distinct school profiles, thereby supporting differentiated and territorially contextualized policy interventions.

In recent years, the field of educational data analytics has expanded significantly, driven by the increasing availability of large-scale educational datasets and advances in computational methods. Educational data analytics seeks to extract meaningful insights from educational data in order to better understand learning outcomes, institutional performance, and the social determinants of education [11,12]. Large-scale assessments such as Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and national examinations like the Brazilian ENEM have become important sources for data-driven research aimed at identifying patterns of educational inequality and institutional performance [1,13]. Within this context, statistical and machine learning techniques—including regression analysis, clustering, and dimensionality reduction methods—have been increasingly used to examine how socioeconomic, demographic, and institutional factors shape educational outcomes. By applying a data-driven analytical framework to ENEM microdata, the present study contributes to this growing body of literature by exploring the determinants of school performance in public schools in the state of Rio de Janeiro.

The dataset used in this study was constructed from the 2023 ENEM microdata provided by the INEP. The original dataset contains individual-level information for students, including exam scores and socioeconomic characteristics. For the purposes of this research, the data were aggregated to the school level, which constitutes the main unit of analysis in the study. Each observation therefore represents a public-school proxy, defined as an aggregation of student-level data belonging to the same school identified in the ENEM database. School-level indicators were calculated by aggregating individual student information, including average exam performance and selected socioeconomic characteristics. After applying the selection criteria and ensuring data completeness, the final analytical dataset consisted of 180 public schools located in the state of Rio de Janeiro.

The case of the state of Rio de Janeiro is of particular empirical relevance. The combination of pronounced socioeconomic inequalities, significant variation across educational networks, and territorial segmentation creates ideal conditions for investigating how structural and contextual factors are associated with the performance of public schools. The state simultaneously hosts high-performing schools and institutions located in peripheral areas facing severe resource constraints, ensuring the variability required for analyses of educational determinants [7].

This study is grounded in the hypothesis that the average performance of public schools in Rio de Janeiro on the 2023 ENEM is significantly influenced by students' socioeconomic context, mediated by demographic and institutional factors. The specific objectives are to: (i) construct a synthetic index that adequately represents the school socioeconomic context; (ii) estimate multiple regression models to identify and quantify the main determinants of performance; (iii) apply rigorous cluster analysis protocols based on consolidated scientific criteria; (iv) conduct robustness tests to assess the stability of the results; and (v) discuss the implications for evidence-based policymaking.

The relevance of this study lies in its methodological and empirical contributions to the understanding of educational inequalities in the context of Rio de Janeiro. The use of advanced statistical techniques—including data transformations, sensitivity analyses, and validation procedures—seeks to ensure the consistency of the estimates and to align the research with established practices in quantitative inquiry. In doing so, the findings provide robust inputs for more effective and context-sensitive public policies, contributing to the pursuit of a more equitable education system.

2. Materials and Methods

This study is characterized as quantitative research of a descriptive and applied nature, focused on the 2023 ENEM microdata made available by the National Institute for Educational Studies and Research Anísio Teixeira (INEP). The investigation adopts a cross-sectional approach, analyzing participants from public schools in the state of Rio de Janeiro, encompassing federal, state, and municipal institutions [1].

The processing of the microdata was conducted using a computational optimization strategy, employing blocks of 1,000,000 records to efficiently manage the data volume. The initial dataset comprised 63,842 participants from public schools in Rio de Janeiro. After applying filters to ensure data completeness—specifically, the presence of valid scores in the four knowledge areas assessed by the ENEM—the final sample consisted of 49,406 students.

To construct the Socioeconomic Status Index, a Principal Component Analysis (PCA) was performed using a set of variables related to the socioeconomic background of students and their families. These variables included indicators commonly associated with educational inequality, such as parental education level, household income, and other socioeconomic characteristics available in the ENEM microdata. The purpose of applying PCA was to reduce the dimensionality of these correlated variables and to synthesize them into a single composite indicator representing the socioeconomic context of the school environment. By combining multiple socioeconomic indicators into a single index, the PCA approach also helped mitigate potential multicollinearity issues in the subsequent regression models (**Appendix A**).

Given that the microdata do not include a unique school identifier, an identifier was constructed by concatenating the municipal code (CO_MUNICIPIO_ESC), the administrative dependency (TP_DEPENDENCIA_ADM_ESC), and a sequential index derived from grouping these variables. This methodological strategy allowed for the differentiation of schools located in the same municipality and under the same administrative dependency, while preserving the necessary analytical granularity and ensuring replicability. Subsequently, the data were aggregated by this identifier, excluding institutions with fewer than 10 participants in the examination. This inclusion criterion aims to ensure minimum statistical representativeness and the stability of the calculated means, resulting in a total of 180 school units.

The variables were organized into four distinct analytical dimensions, grounded in the literature on the determinants of educational performance. The dependent variable—academic performance—corresponds to the arithmetic mean of the scores obtained in the four objective areas of the ENEM: Natural Sciences, Languages and Codes, Human Sciences, and Mathematics. This synthetic measure provides a comprehensive indicator of school performance, capturing diverse competencies and minimizing biases associated with specific knowledge domains. The independent variables were categorized into three dimensions: (i) Socioeconomic dimension—parental education, calculated as the mean of responses to questions Q001 and Q002 of the socioeconomic questionnaire, measured on an ordinal scale from 1 to 8, and household income, corresponding to question Q006, measured on an ordinal scale from 0 to 16; (ii) Demographic dimension—the proportion of female students and the proportion of students self-identified as non-white (Black, Brown/mixed-race, Indigenous, or Asian); and (iii) Institutional dimension—the administrative dependency of the school, operationalized through dummy variables for Federal School and Municipal School, with State School serving as the reference category.

Exploratory analysis revealed a high correlation between parental education and household income (r = 0.843), indicating a potential multicollinearity problem that could compromise the stability of the estimates. To mitigate this limitation, Principal Component Analysis (PCA) was applied, a well-established statistical technique for dimensionality reduction and the synthesis of correlated variables. Bartlett's test of sphericity [14] confirmed the suitability of the data for PCA (χ^2 = 211; p < 0.001), rejecting the null hypothesis that the correlation matrix is equal to the identity matrix. The first principal component explained 92.19% of the joint variance of the two variables, with symmetric loadings of 0.707 for both. This component was interpreted as a Socioeconomic Status Index (SEI), replacing the two original variables in the model and effectively eliminating multicollinearity. An analysis of Variance Inflation Factors (VIF) confirmed that all values remained below 2.5 after the transformation, meeting established criteria for statistical adequacy.

To further assess potential multicollinearity among the explanatory variables, Variance Inflation Factors (VIF) were calculated for all predictors in the regression model. The results indicated elevated VIF values for household income (6.4) and parental education (4.7), confirming the presence of relevant multicollinearity between these two variables. Given this overlap of information, Principal Component Analysis (PCA) was employed to combine these variables into a single SEI. After this transformation, all VIF values in the final regression model remained below 2.5, indicating that multicollinearity was no longer a concern.

Principal Component Analysis (PCA) was applied to the variables parental education and household income in order to construct the Socioeconomic Status Index. Because the analysis involved only two highly correlated variables, two components were initially generated. The first principal component (PC1) explained 92.19% of the total variance, while the second component accounted for the remaining 7.81%. Given its dominant explanatory power, only the first component was retained and interpreted as the Socioeconomic Status Index used in the regression models.

The first stage of the analysis consisted of estimating a multiple linear regression model with robust HC3 standard errors [15], a specification that provides protection against heteroskedasticity and ensures valid statistical inference even in the presence of non-constant error variances:

$$\begin{aligned}\text{Performance} = \beta_0 + \beta_1(\text{SEI}) + \beta_2(\text{prop\_non\_white}) + \beta_3(\text{prop\_female})\\ + \beta_4(\text{federal\_school}) + \beta_5(\text{municipal\_school}) + \varepsilon\end{aligned}$$

The specification of the regression model followed a theoretically guided variable selection strategy grounded in the literature on the determinants of educational performance. The dependent variable corresponds to the average score obtained by each school proxy in the four objective areas of the ENEM examination. The explanatory variables were organized into three analytical dimensions: socioeconomic, demographic, and institutional. Socioeconomic context was represented by the SEI, constructed through Principal Component Analysis from parental education and household income. Demographic controls included the proportion of female students and the proportion of students self-identified as non-white. Institutional characteristics were captured through dummy variables representing federal and municipal schools, with state schools serving as the reference category. Variables were retained in the model based on theoretical relevance and empirical evidence in the literature on educational inequality, rather than purely statistical stepwise procedures, in order to avoid model instability and omitted variable bias. Model assumptions were rigorously assessed through a comprehensive battery of diagnostic tests: the Jarque–Bera test [16] for residual normality, the Breusch–Pagan test [17] for homoskedasticity, and an analysis of Cook's Distance [18] to identify influential observations.

To verify the stability and robustness of the results, two complementary validation strategies were employed. First, a log-linear model was estimated, in which a logarithmic transformation was applied to the dependent variable, altering the functional form of the relationship between variables and allowing the coefficients to be interpreted in percentage terms. This alternative specification aims to assess whether the results hold under a different mathematical structure, thereby enhancing the reliability of the conclusions. Additionally, a bootstrap procedure with 1,000 resamples was implemented to nonparametrically validate the stability of the estimated coefficients. The bootstrap provides a robust method for assessing the sampling distribution of the estimates without imposing specific distributional assumptions, yielding empirical confidence intervals and testing the sensitivity of the results to variations in sample composition.

Complementing the regression modeling, a cluster analysis was conducted using the K-Means algorithm [19], configured with 10 random initializations, a maximum of 300 iterations, and an early stopping criterion for improvements below 0.001. The initial selection of centroids sought to ensure better distribution and accelerate algorithm convergence. The determination of the optimal number of clusters was based on three integrated and scientifically validated criteria: (i) internal quality, assessed using the Silhouette Coefficient [20] and the Calinski–Harabasz Index [21]; (ii) stability, measured through 50 bootstrap resamples using 80% of the data and quantified by the Adjusted Rand Index (ARI) [22]; and (iii) parsimony, prioritizing the smallest k among those exhibiting the highest Silhouette Coefficient and greatest stability, considering a technical tie for differences smaller than 0.01. Candidate solutions for different values of k were evaluated and compared using these criteria. The silhouette coefficient was used to assess cluster cohesion and separation, the Calinski–Harabasz index to evaluate the ratio of between-cluster to within-cluster variance, and the Adjusted Rand Index (ARI) to measure the stability of the clustering solution across bootstrap resamples. The results showed convergence of these indicators for $k = 4$, which presented a favorable balance between internal quality, stability, and parsimony. Therefore, the four-cluster solution was selected as the optimal segmentation of the dataset.

All analyses were conducted in Python 3.11, using specialized libraries and rigorously adhering to best practices in scientific programming. The study relied exclusively on public, aggregated, and anonymized data, in compliance with ethical research guidelines and current data protection regulations. Aggregation at the school level eliminates any possibility of individual identification, ensuring participant anonymity. The adopted methodology integrates established statistical procedures with analytical innovations, seeking to balance scientific rigor and practical applicability. The combination of dimensionality reduction techniques, robust modeling, and multiple validation strategies aims to ensure the reliability of the results and their suitability for informing evidence-based public policies.

The use of large-scale educational microdata also requires careful attention to data governance and ethical considerations. In this study, all analyses were conducted using publicly available and anonymized datasets provided by the INEP. The data do not contain personally identifiable information, and the aggregation of observations at the school level further eliminates any possibility of individual identification. These procedures ensure compliance with ethical research standards and current data protection regulations while allowing the responsible use of large-scale educational data for scientific research.

To facilitate understanding of the methodological architecture and to highlight the logical sequence of the analyses performed, **Figure 1** presents a concise flowchart mapping the entire investigative process. The diagram illustrates the stages from the initial loading and preprocessing of ENEM microdata to the generation of the final results, highlighting key methodological decision points, the statistical techniques employed, and the convergence between regression and clustering approaches. This visual representation enables readers to follow the analytical workflow and to understand how each procedure contributes to addressing the proposed objectives, demonstrating the integration of the methodological stages.

3. Results

The descriptive analysis revealed that the 180 public schools in the state of Rio de Janeiro exhibited an average performance of 533.5 points on the 2023 ENEM, with a standard deviation of 40.8 points. The distribution was approximately normal, with a range of 208 points between the lowest-performing school (462.8) and the highest-performing school (670.9). This substantial variability highlights the heterogeneity of the public education system in the state, with institutions operating at markedly different quality levels. **Figure 2** presents a boxplot of the distribution of average performance, illustrating the dispersion of the data, the presence of outliers, and a slight right-skewness in the performance of the schools analyzed.

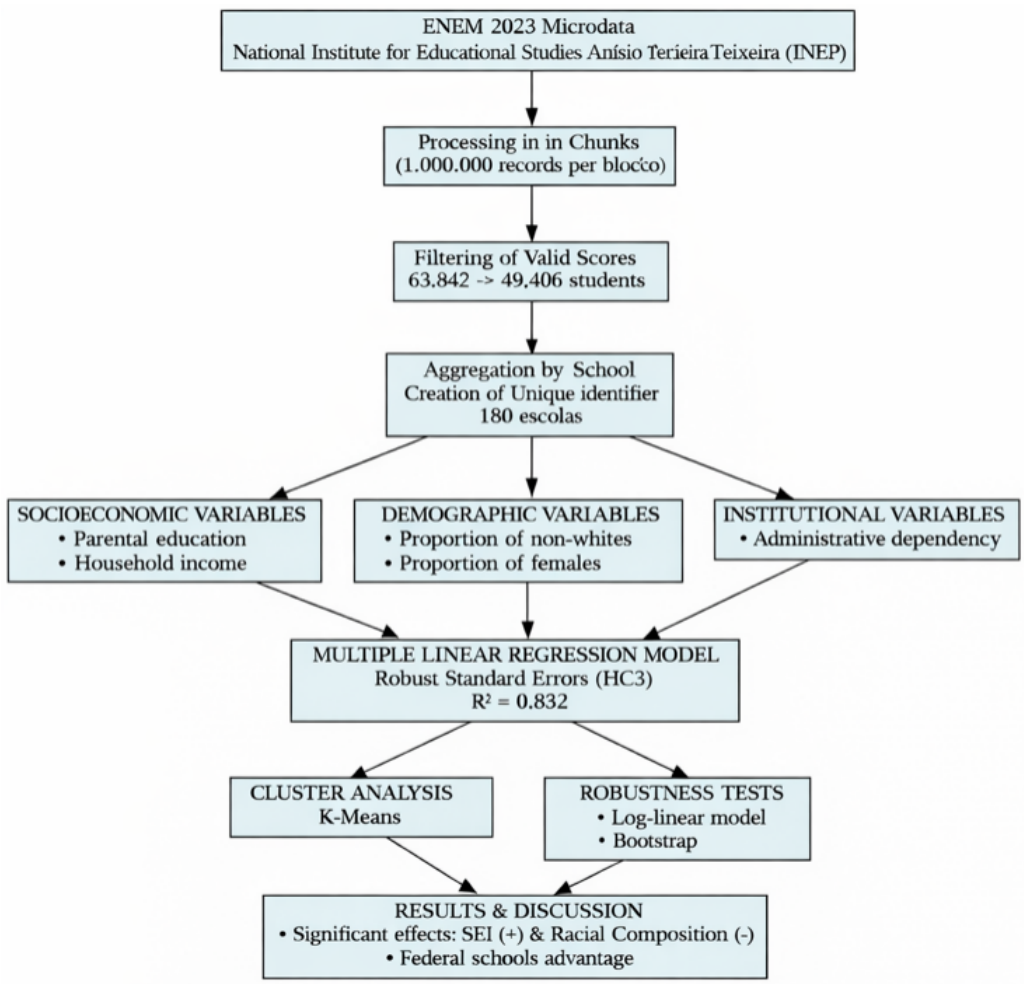

Figure 1. Research Methodological Flowchart.

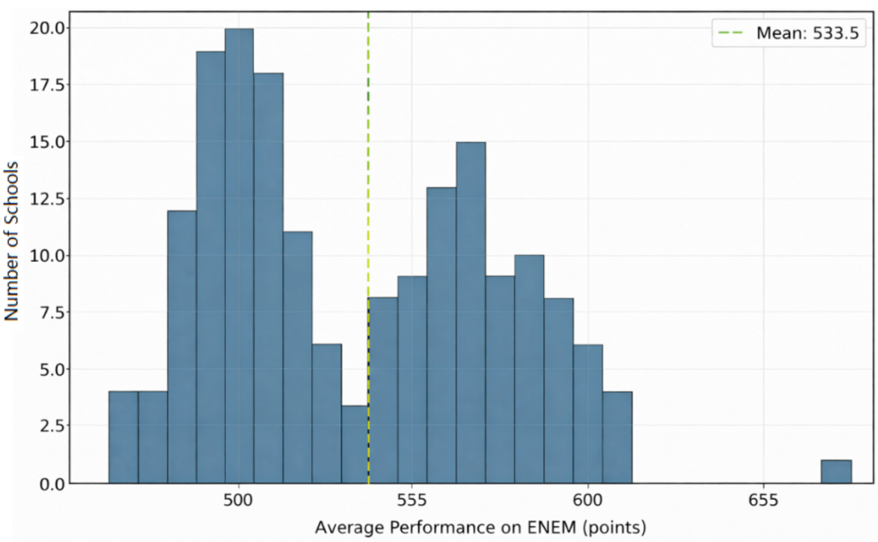

Figure 2. Distribution of Average School Performance.

The characterization of explanatory variables revealed patterns consistent with the literature on educational inequalities in Brazil. The average level of education of parents was 5.0 points on a scale of 1 to 8, corresponding approximately to complete high school education. The average family income reached 4.2 points on a scale of 0 to 16, equivalent to approximately three minimum wages. These indicators reflect the typical profile of families, traditionally marked by very limited financial resources and median parental education.

Demographically, the sample composition showed a predominance of self-declared non-white students (65.3%) and a slight female majority (56.1%). This racial distribution reflects the demographic composition of the state of Rio de Janeiro and suggests that the public education system predominantly serves historically marginalized populations. Institutionally, state schools predominated (67.2%), followed by municipal (24.4%) and federal (8.4%) schools.

The variables of parental education and family income showed a high correlation with school performance (0.757 and 0.836, respectively), justifying the construction of a synthetic index. Bartlett's sphericity test confirmed the adequacy of the data ($\chi^2 = 1527.4$; $p < 0.001$), rejecting the null hypothesis of independence between the variables and validating the technical application of PCA (**Appendix B**).

The multiple linear regression showed an adjusted R^2 of 0.832. This result indicates that the model is capable of explaining more than four-fifths of the variation in the average performance of the schools analyzed, demonstrating high explanatory power. The overall significance of the equation is confirmed by the F statistic (173.88; $p < 0.001$) and by the AIC (1,531) and BIC (1,550) information criteria, according **Table 1**.

Table 1. Multiple Linear Regression Results with HC3 Robust Standard Errors.

Variable	Coefficient	Standard Error	Z	*p*-Value	95% CI
Socioeconomic Status Index (SEI)	18.87	1.30	14.52	<0.001	[16.32; 21.42]
Proportion of Non-White Students	−68.48	13.46	−5.09	<0.001	[−94.86; −42.11]
Proportion of Female Students	3.93	21.59	0.18	0.855	[−38.39; 46.25]
Federal School	39.46	4.12	9.59	<0.001	[31.40; 47.53]
Municipal School	6.60	12.96	0.51	0.611	[−18.80; 32.00]
Constant	556.22	12.55	44.32	<0.001	[531.62; 580.82]

Three variables exhibited statistical significance at the 1% level, consolidating their role as determinants of educational performance. The Socioeconomic Status Index (SEI) displayed a positive coefficient of 18.87 points ($p < 0.001$), indicating that a one–standard deviation increase in the socioeconomic context is associated with an average increase of 18.87 points in school performance. This result is consistent, who emphasize the importance of family background and the average socioeconomic level of schools [4,8].

The proportion of non-white students showed a negative coefficient of −68.49 points ($p < 0.001$), a result that warrants careful interpretation. It is important to note that racial composition and socioeconomic conditions are often structurally interconnected in the Brazilian context. Historically marginalized racial groups tend to face greater socioeconomic disadvantages, including lower household income, lower parental education levels, and reduced access to educational resources. However, in the present study the effect associated with the proportion of non-white students remains statistically significant even after controlling for socioeconomic status through the Socioeconomic Status Index (SEI). This result suggests that racial disparities in educational outcomes may reflect broader structural inequalities, including differences in school environments, expectations, and access to opportunities, which extend beyond purely economic factors. Even after controlling for socioeconomic and institutional variables, schools with higher shares of Black, Brown (mixed-race), and Indigenous students tend to achieve lower outcomes. This finding highlights the persistence of racial inequalities that extend beyond material conditions, in line with concept of cultural capital and with the evidence provided regarding racial disparities in the Brazilian education system [6,7].

Federal schools exhibited an average performance 39.46 points higher than that of state schools ($p < 0.001$), suggesting a strong association between administrative dependency and educational outcomes. This difference may reflect structural and institutional characteristics of the federal education network in Brazil. Federal schools—particularly Federal Institutes and Colégios de Aplicação—generally benefit from higher levels of public investment per student, more selective admission processes, and more stable teaching staff with higher levels of academic qualification. In addition, these institutions often operate with stronger organizational autonomy, better infrastructure, and pedagogical models that integrate academic and technical education. These institutional features may contribute to more favorable learning environments and help explain their consistently higher performance in national assessments such as the ENEM. This result resonates with by underscoring the role of institutional arrangements in shaping educational quality, and with human capital theory, which posits that sustained investments in organizational structures and pedagogical practices enhance educational returns [2,5]. In contrast, the proportion of female students and municipal dependency were not statistically significant (p = 0.611 and p = 0.855, respectively), but were retained in the model as control variables, thereby ensuring specification robustness and minimizing the risk of omitted variable bias.

Regression assumptions were assessed through a comprehensive set of diagnostic tests. The Jarque–Bera test (statistic = 3.98; p = 0.137) and the Shapiro–Wilk test (W = 0.99; p = 0.093) [23], did not reject the normality of the residuals, while the Breusch–Pagan test (χ^2 = 6.07; p = 0.299) indicated no evidence of heteroskedasticity. Nonetheless, HC3 robust standard errors were employed to enhance estimation reliability. Cook's Distance identified 17 influential schools, all below the critical threshold (4/n ≈ 0.022); therefore, these observations were retained to preserve the representativeness of the public school network. Their exclusion would not substantially alter the estimated coefficients (**Appendix C**).

As a complementary robustness check, a log-linear regression was estimated with the dependent variable transformed using the natural logarithm. This specification is appropriate given that the outcome variable is continuous, positive, and mildly skewed, allowing coefficients to be interpreted in percentage terms. The results preserved both the signs and statistical significance of the linear model, with a similar adjusted R^2 (83.3%), confirming the robustness of the findings across alternative functional forms and reinforcing the credibility of the conclusions (**Appendix C**).

Finally, bootstrap validation with 1,000 resamples was employed to assess the stability of the coefficients under variations in sample composition. The main effects maintained consistent direction and statistical significance, indicating that the results are not driven by specific observations and further reinforcing the empirical robustness of the estimated model (**Appendix C**).

4. Discussion

In the second stage of the study, the determination of the optimal number of clusters was based on multiple and convergent criteria: the silhouette coefficient (0.474), inertia from the elbow method (288.27), the Adjusted Rand Index (ARI) for stability (0.962), and the Calinski–Harabasz index (125.23). These indicators converged on a four-cluster solution (k = 4), a configuration that simultaneously optimizes internal quality, stability, and parsimony, according to **Figure 3**.

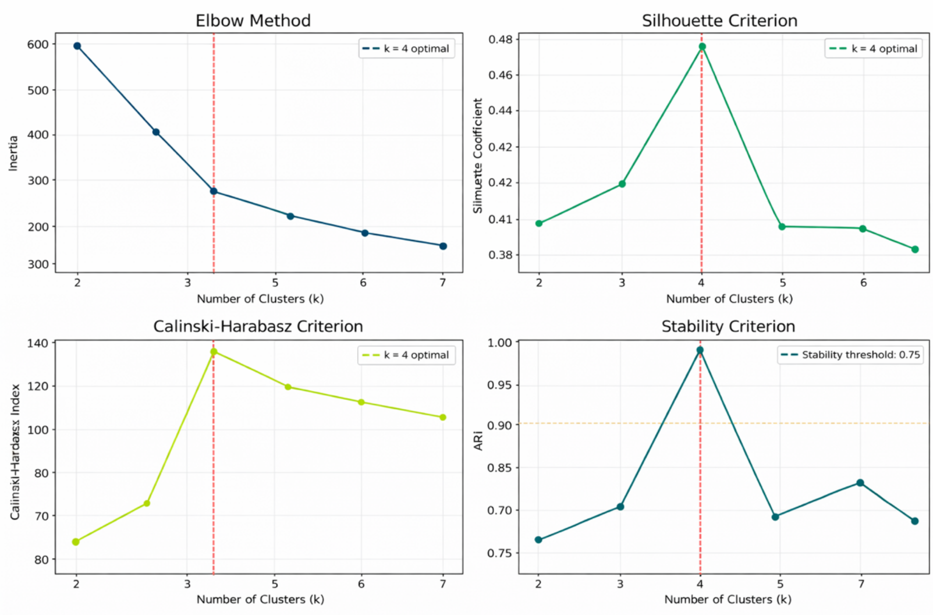

Figure 3. Determination of the Optimal Number of Clusters.

The stability of the clusters was rigorously evaluated using bootstrap resampling with 100 iterations. The Adjusted Rand Index (ARI) distribution had a mean of 0.966, with values concentrated above 0.94, according **Figure 4**. In all iterations, the ARI exceeded the threshold of 0.75 recommended, evidencing high reproducibility of the segmentation and confirming consistent patterns of the clusters [24]. Overall, the cluster solution reveals four distinct institutional profiles within the public education system of Rio de Janeiro. The first profile corresponds to high-performing federal schools characterized by favorable socioeconomic conditions and lower proportions of non-white students. The second profile represents the largest group of state schools, marked by lower socioeconomic status and lower academic performance. The third profile includes socioeconomically advantaged segments of the state and municipal networks that achieve performance levels comparable to federal institutions. Finally, the fourth profile corresponds to a small group of municipal schools with moderate performance and higher socioeconomic vulnerability. Together, these profiles illustrate how combinations of socioeconomic context, racial composition, and institutional arrangements are associated with distinct educational outcomes.

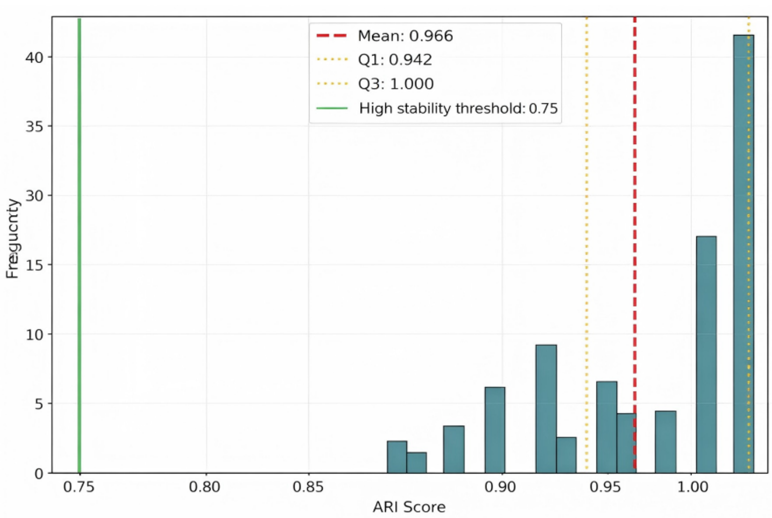

Figure 4. Distribution of the Adjusted Rand Index (Stability Criterion).

The segmentation resulted in the identification of four distinct profiles of public schools, characterized by combinations of socioeconomic, demographic, and institutional attributes (**Appendix D**).

Cluster 0 (n = 27; 15.0%): grouped exclusively federal schools. It had an average performance of 573.4 points, a positive ISE of 0.34, a proportion of non-white students of 0.45, and gender balance (0.48 female students). This is the group with the highest income and homogeneity.

Cluster 1 (n = 94; 52.2%): comprises the majority of the sample, concentrating state schools. With an average performance of 501.3 points and an ISE of −1.00, this is the group with the lowest performance and highest socioeconomic vulnerability. The proportion of non-white students was 0.54 and that of female students was 0.59.

Cluster 2 (n = 55; 30.6%): stood out for its higher average SES (1.61) and academic performance comparable to that of federal schools, achieving 570.5 points on the exam. With the lowest relative presence of non-white students (0.30) and an intermediate female proportion (0.52), this group represents the most socioeconomically favored segment of the state and municipal school system.

Cluster 3 (n = 4; 2.2%): corresponds to a low-representative group, consisting exclusively of municipal schools. It had an average performance of 512.2 points, an ISE of −0.94, an intermediate proportion of non-white students (0.51), and the highest female presence among the clusters (0.60). **Table 2** shows the data.

Table 2. Average Profiles of the Identified Clusters.

Characteristics	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Average Performance	573.35	501.27	570.51	512.23
Mean SEI	0.34	−1.00	1.61	−0.93

Table 2. *Cont.*

Characteristics	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Proportion of Non-White Students	0.44	0.54	0.29	0.51
Proportion of Female Students	0.48	0.59	0.53	0.52
Federal School	1.00	0.00	0.00	0.00
Municipal School	0.00	0.00	0.00	1.00
Performance Profile	High	Low	High	Low
Number of Schools	27	94	55	4

The characterization of the clusters is based on the mean centroids of the segmentation variables, which synthesize the typical structural profiles of each group. This approach enables direct comparison across different school contexts, revealing how specific combinations of socioeconomic, demographic, and institutional factors are associated with distinct levels of educational performance. **Figure 5**, presented below, details these average profiles, visually reinforcing the distinctions among the identified clusters in terms of socioeconomic status, racial composition, average performance, and relative size.

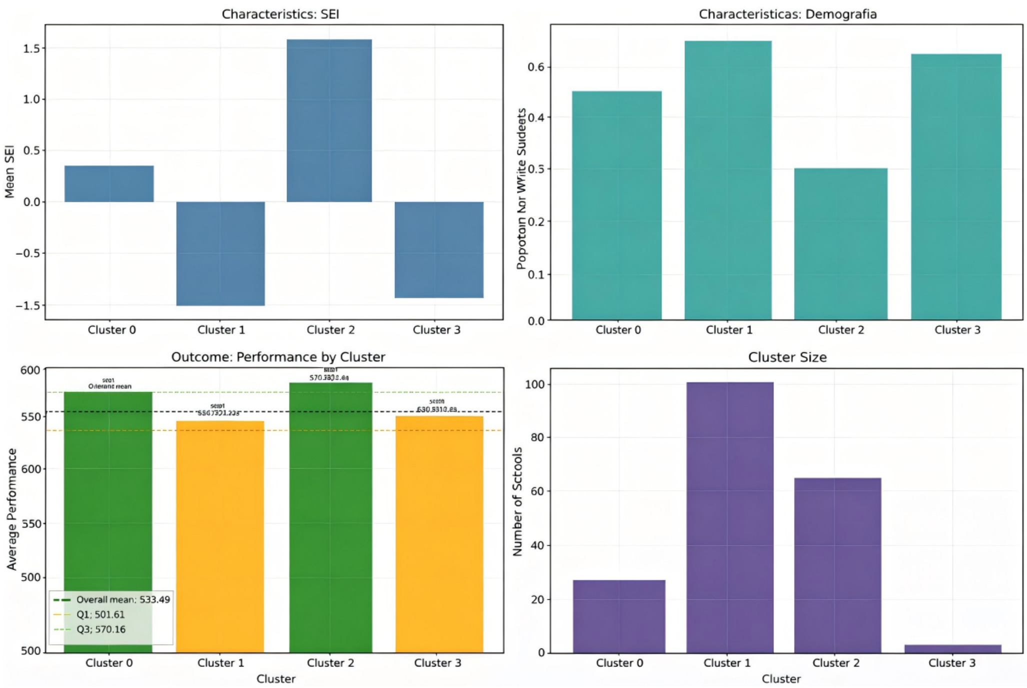

Figure 5. Average Cluster Profiles.

The magnitude of the estimated effects has immediate relevance for educational policy. The positive coefficient of the Socioeconomic Status Index (18.87 points per standard deviation) indicates that improvements in the socioeconomic context are associated with higher performance, whereas the negative coefficient for the proportion of non-white students (−68.49 points) reveals racial inequalities that extend beyond material conditions. The concentration of federal schools within a high-performing cluster, with parallels observed among segments of state schools, suggests that institutional and pedagogical arrangements may partially offset financial constraints. The consistency of the results and the stability of the clustering solution, who emphasized the role of the ENEM in exposing territorial inequalities, and with the emphasis on the centrality of schools' average socioeconomic level. Taken together, the evidence demonstrates that educational asymmetries in the state are structural in nature and require targeted policy interventions to be effectively addressed [3,9].

It is also important to acknowledge that ENEM scores may be influenced by additional factors not directly captured in the present model. Potential confounding variables include differences in school infrastructure, teacher qualifications, pedagogical practices, and access to educational resources. Moreover, contextual elements such as neighborhood conditions, school management practices, and students' prior educational trajectories may also influence academic outcomes. Although the present study controls for key socioeconomic, demographic, and institutional characteristics, the absence of these additional variables suggests that the estimated associations should be interpreted with caution. Future research integrating data from the School Census and other educational databases could provide a more comprehensive understanding of the determinants of school performance.

5. Conclusions

The study demonstrated that the performance of public schools in the state of Rio de Janeiro on the 2023 ENEM is largely shaped by structural inequalities of a socioeconomic, institutional, and racial nature. The analysis showed that these dimensions interact in a persistent manner, giving rise to distinct institutional profiles and revealing the heterogeneity of the state's public education system.

The findings point to the need for targeted educational policies. Schools with a higher proportion of non-white students require equity-oriented programs that combine pedagogical support, adequate infrastructure, and anti-racist teacher training. The experience of federal schools indicates that consistent organizational and pedagogical models can be replicated across other education networks, contributing to improvements in teaching quality. The typology developed in this study provides practical guidance for the strategic allocation of resources and for interventions tailored to the identified vulnerabilities.

The robustness of the findings was further assessed through complementary sensitivity analyses. These procedures included the estimation of an alternative log-linear specification of the regression model, the identification and exclusion of influential observations, and bootstrap resampling with 1,000 iterations. The consistency of the estimated coefficients across these different specifications indicates that the main conclusions of the study are not driven by specific model assumptions or sample configurations.

The study's limitations highlight relevant avenues for future research. The use of cross-sectional data restricts causal inference, underscoring the importance of longitudinal analyses that allow for tracking the evolution of inequalities over time. Moreover, further improvements could be achieved through integration with additional data sources. Variables from the School Census, such as infrastructure, teacher profiles, and management practices, would enable a deeper understanding of pedagogical processes, while indicators from SAEB would allow for temporal analyses of performance. The focus on the state of Rio de Janeiro also limits the generalizability of the findings, suggesting the need for interstate comparisons. Another relevant dimension concerns the potential influence of spatial and regional factors within the state of Rio de Janeiro. The state is characterized by pronounced territorial inequalities, including differences between metropolitan areas, mid-sized municipalities, and peripheral regions. Variations in local economic conditions, urban infrastructure, and access to educational resources may contribute to heterogeneous school performance across territories. Although the present study focuses primarily on socioeconomic, demographic, and institutional variables, spatial context may also shape educational outcomes through neighborhood effects and local policy environments. Future research could therefore incorporate spatial analysis techniques or regional controls to further explore how territorial dynamics interact with school performance. An additional limitation concerns the cross-sectional nature of the dataset, which is restricted to a single edition of the ENEM examination. Although this approach allows the identification of structural associations between socioeconomic, demographic, and institutional factors and school performance, it does not capture the temporal dynamics of educational inequalities. Longitudinal analyses using multiple years of ENEM microdata could provide valuable insights into how these patterns evolve over time and whether policy interventions are capable of reducing performance disparities across schools. Finally, the relative efficiency of schools was not assessed; future studies could therefore employ frontier methods, such as Data Envelopment Analysis (DEA), to complement mean-based analytical approaches.

The findings point to the need for targeted educational policies [25,26]. Schools with a higher proportion of non-white students require equity-oriented programs that combine pedagogical support, adequate infrastructure, and anti-racist teacher training. The experience of federal schools indicates that consistent organizational and pedagogical models can be replicated across other education networks, contributing to improvements in teaching quality. The typology developed in this study provides practical guidance for the strategic allocation of resources and for interventions tailored to the identified vulnerabilities.

The study's limitations highlight relevant avenues for future research. The use of cross-sectional data restricts causal inference [27–29], underscoring the importance of longitudinal analyses that allow for tracking the evolution of inequalities over time. Moreover, further improvements could be achieved through integration with additional data sources. Variables from the School Census, such as infrastructure, teacher profiles, and management practices, would enable a deeper understanding of pedagogical processes, while indicators from SAEB would allow for temporal analyses of performance. The focus on the state of Rio de Janeiro also limits the generalizability of the findings, suggesting the need for interstate comparisons. Finally, the relative efficiency of schools was not assessed; future studies could therefore employ frontier methods, such as Data Envelopment Analysis (DEA), to complement mean-based analytical approaches.

In conclusion, the construction of a more equitable education system in Rio de Janeiro requires evidence-based public policies aimed at disrupting the historical cycle of inequalities that constrain the opportunities of thousands of students in the state. This study contributes by providing a robust quantitative diagnosis capable of informing targeted interventions and supporting strategic decision-making in the field of educational policy, while also opening new avenues for academic research focused on school efficiency, the temporal dynamics of inequalities, and comparative analyses across regional contexts. Although this study focuses on the state of Rio de Janeiro, the analytical framework developed here can be applied to other educational systems and regional contexts. The combination of dimensionality reduction techniques, regression modeling, and cluster analysis provides a flexible methodological approach for identifying structural determinants of school performance. Applying this framework to other Brazilian states or to international educational datasets could enable comparative analyses and contribute to a broader understanding of educational inequalities across different institutional environments.

While the present study relies on established statistical techniques such as Principal Component Analysis, multiple linear regression, and clustering methods, recent advances in machine learning offer additional possibilities for educational data analysis. Approaches such as random forests, gradient boosting models, and neural networks may capture complex nonlinear relationships and interactions among variables that are not easily detected by traditional econometric models. Although the primary objective of this study was to identify interpretable structural determinants of school performance, future research could explore these machine learning techniques to complement regression-based approaches and further investigate patterns in large-scale educational datasets such as ENEM microdata.

Author Contributions

Conceptualization, J.V.G.R., R.A.C., R.F.F., and F.P.P.; methodology, J.V.G.R. and F.P.P.; validation, J.V.G.R., R.A.C., R.F.F., and F.P.P.; formal analysis, J.V.G.R., R.A.C., R.F.F., and F.P.P.; investigation, J.V.G.R. and F.P.P.; data curation, J.V.G.R., R.A.C., R.F.F., and F.P.P.; writing—original draft preparation, J.V.G.R., R.A.C., R.F.F., and F.P.P.; writing—review and editing, R.F.F. and F.P.P. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by Fundação para a Ciência e a Tecnologia (grant UID/05064/2025 - https://doi.org/10.54499/UID/05064/2025) and ALT2030-FSE+-01459600.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Research data can be found in https://github.com/joaovgr-pixel/Determinantes-do-desempenho-das-Escolas-Publicas-do-RJ-no-ENEM-2023-/tree/main.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

The authors used ChatGPT (OpenAI, GPT-5.3) solely for grammar checking, sentence structure refinement, and improving the readability of the English text in this manuscript. The authors take full responsibility for all academic content, including all ideas, data, analyses, and conclusions presented herein. The use of AI was thoroughly reviewed and supervised by the authors.

Appendix A

The initial analysis examined the degree of linear association among the independent variables using the Pearson correlation matrix (**Figure A1**). The results revealed a high correlation (r = 0.843) between parents' educational attainment and household income, exceeding the 0.70 threshold adopted as an indicator of problematic collinearity. This overlap of information violates the assumptions of multiple linear regression and may distort the interpretation of the estimated coefficients.

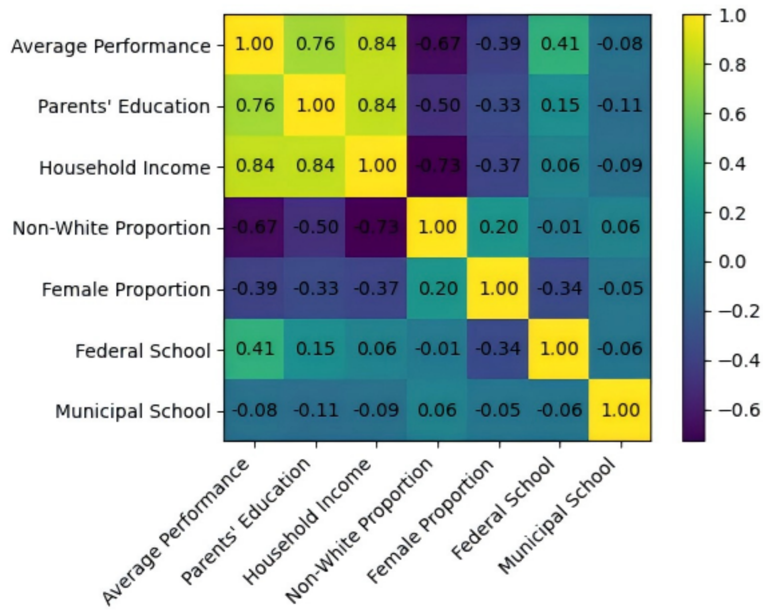

Figure A1. Correlation matrix between study variables.

To complete the diagnosis, the Variance Inflation Factor (VIF) was calculated for all variables in the model (**Table A1**). The average VIF was 2.06, with maximum values close to 7, confirming that the relevant multicollinearity was concentrated in the pair formed by parental education and family income. This evidence justified the application of Principal Component Analysis (PCA) to condense these variables into a single synthetic index, reducing redundancy without substantial loss of information.

Table A1. Variance Inflation Factors.

Variable	VIF	Interpretation
Household Income	6.4	Severe multicollinearity
Parents' Education	4.7	Close to the threshold of concern
Non-White Proportion	2.1	No evidence of multicollinearity
Female Proportion	1.9	No evidence of multicollinearity
Federal School	1.2	No evidence of multicollinearity
Municipal School	1.1	No evidence of multicollinearity

Appendix B

Principal Component Analysis was employed to construct the Socioeconomic Status Index (SEI) from the variables parents' educational attainment and household income. Bartlett's test of sphericity yielded a value of χ^2 =

1,527.4 ($p < 0.001$), rejecting the null hypothesis of an identity matrix and confirming the suitability of the procedure.

Figure A2 presents the variance explained by each component, highlighting that the first component (PC1) retained 92.19% of the total variance, indicating a high level of data reduction and synthesis. As illustrated in **Figure A3**, the loadings were symmetric, indicating equivalent contributions of parents' education and household income to the construction of the index. These results confirm the statistical robustness of the SEI as a synthetic measure of the school socioeconomic context, while also eliminating the multicollinearity identified in **Appendix A**.

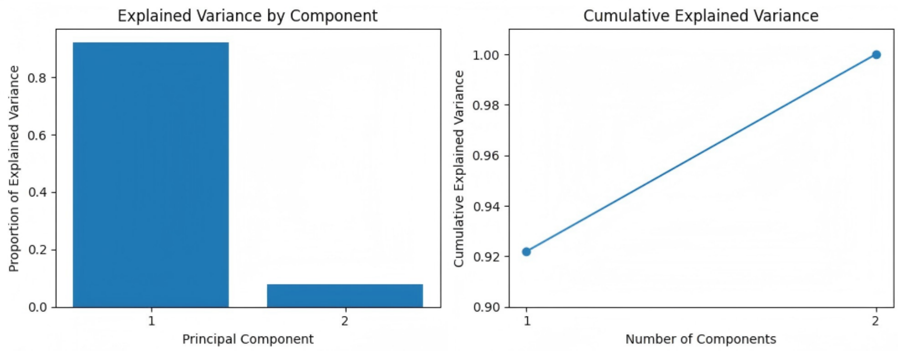

Figure A2. Variance explained by principal components.

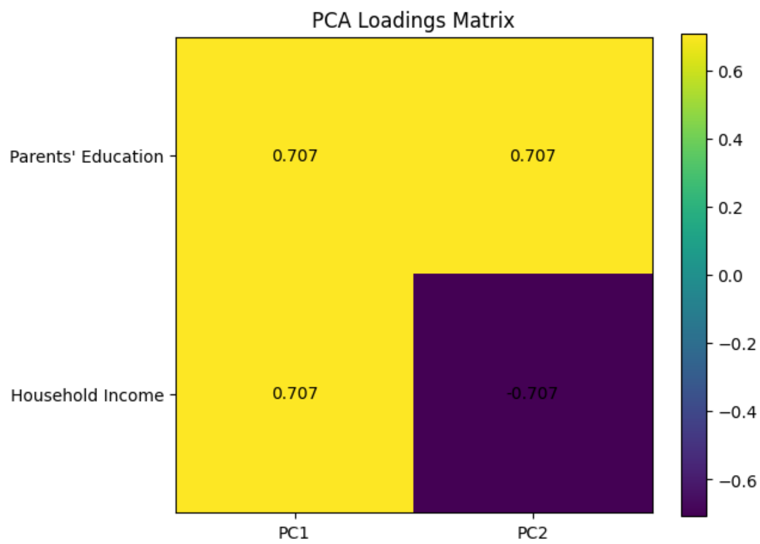

Figure A3. Principal Component Loadings.

Figure A2 shows the variance explained by each component, highlighting that the first (PC1) retained 92.19% of the total variance, which shows a high capacity for synthesis. As illustrated in **Figure A3**, the loadings were symmetrical, indicating an equivalent contribution of parental education and family income to the composition of the index. These results confirm the statistical robustness of the ISE as a synthetic socioeconomic measure.

Appendix C

Statistical consistency was assessed through diagnostic tests and sensitivity analyses to ensure that the conclusions were not influenced by violations of the assumptions of linear regression. The inspection of residuals, presented in **Figure A4**, evaluates linearity, homoscedasticity, and normality.

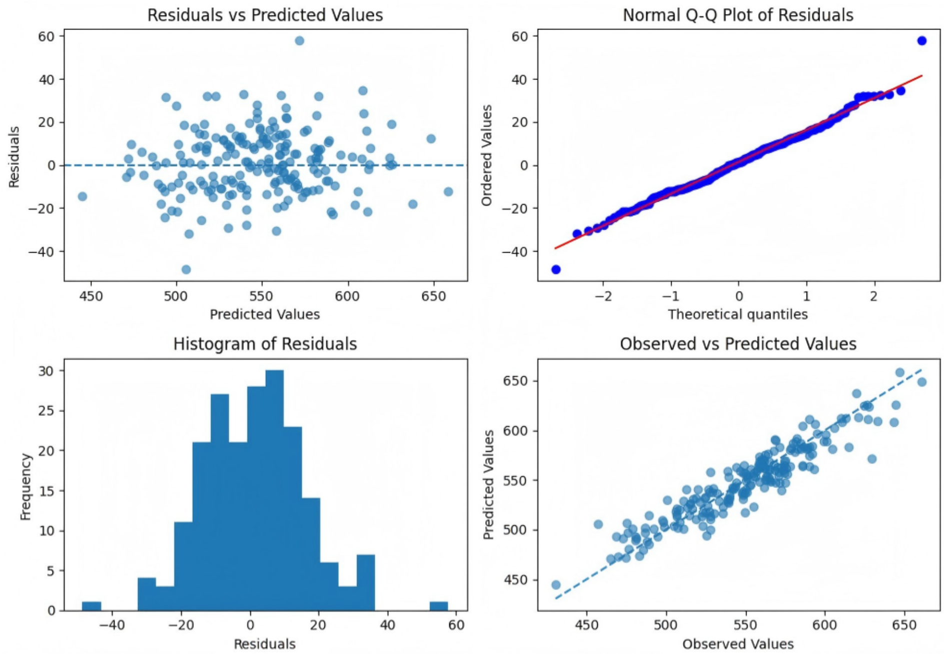

Figure A4. Visual diagnostics of the regression model residuals.

The upper left panel (residuals and predictive values) revealed a dispersion pattern without pronounced curvatures, but with evidence of non-constant variances, suggesting potential heteroscedasticity. The Q-Q plot (upper right panel) indicated adherence to the normal distribution. The lower right panel, which relates residuals and adjusted values, did not show systematic patterns that compromised linearity, although it reiterated the need to address heteroscedasticity. This condition was confirmed by the Breusch-Pagan test ($p < 0.05$), which led to the adoption of MacKinnon and White's robust covariance estimator H_3, capable of adjusting standard errors without altering the estimated coefficients, thereby increasing the reliability of the inferences.

Additionally, influential observations were identified using Cook's distance, shown in **Figure A5**. The conventional criterion of 4/n (n = 180) set the influence threshold at 0.0222. Seventeen schools (9.4% of the sample) were identified above this critical value, characterized by extreme socioeconomic or academic performance profiles.

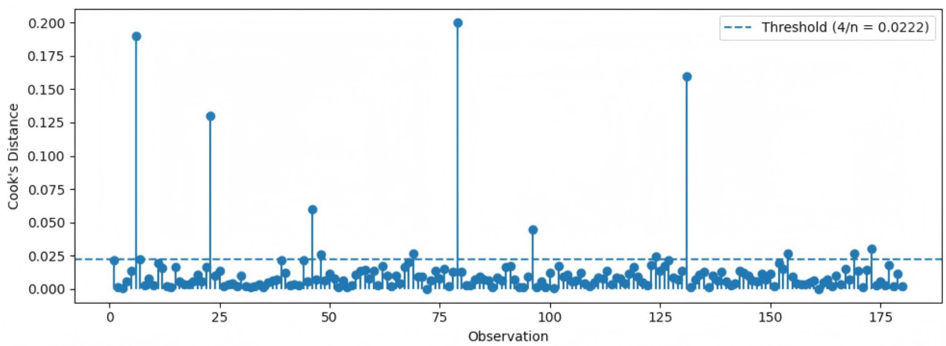

Figure A5. Cook's distance for identifying influential observations.

To assess the impact of these units on the results, the model was re-estimated excluding them, and the resulting coefficients were compared with those from the full specification (**Table A2**). The differences were marginal for most variables, preserving the magnitude, direction, and statistical significance of the effects. The largest variation was observed for the "Municipal School" variable (7.57 points), indicating localized sensitivity without substantively altering the conclusions. As can be seen in the following section.

Table A2. Sensitivity Analysis: Exclusion of Influential Observations.

Variable	Original Coef.	Coef. without Influential Observations	Difference
Constant	556.22	556.31	0.09
SEI	18.87	18.56	0.31
Non-White Proportion	−68.49	−70.55	2.06
Female Proportion	3.93	4.80	0.87
Federal School	39.46	38.96	0.50
Municipal School	6.60	14.17	7.57

Therefore, although heteroskedasticity is present—hence the adoption of a robust estimator—no violations were detected that would compromise the validity of the coefficients.

As an additional step, a log-linear specification was estimated, in which the natural logarithm of average performance was used as the dependent variable. This transformation allows the coefficients to be interpreted as elasticities or percentage effects. The goodness-of-fit indicators reported in **Table A3** show that the log-linear model performed slightly better, with a marginally higher adjusted R^2 compared to the initial model, as well as lower AIC and BIC values. However, this advantage should be interpreted with caution, since the dependent variables are expressed on different scales.

Table A3. Comparison between Linear and Log-Linear Models.

Model	R^2	Adjusted R^2	AIC	BIC	F Statistic
Linear	0.836	0.832	1,530.97	1,550.13	173.88
Log-Linear	0.838	0.833	−734.94	−715.78	191.03

The robustness of the estimates was verified using bootstrap resampling. **Table A4** presents the estimated coefficients for the log-linear model. The results confirm the main findings: ISE maintains a positive and statistically significant effect (0.035), the proportion of non-white students shows a negative association (−0.128), and federal schools exhibit a performance advantage (0.074). Both the signs and the statistical significance of these coefficients are consistent with the linear model, reinforcing the robustness of the conclusions.

Table A4. Log-Linear Model Coefficients.

Variable	Coefficient	Standard Error	*p*-Value	Significance
Constant	6.315	0.022	<0.001	***
SEI	0.035	0.002	<0.001	***
Non-White Proportion	−0.128	0.025	<0.001	***
Female Proportion	0.016	0.036	0.664	
Federal School	0.074	0.007	<0.001	***
Municipal School	0.012	0.026	0.634	

Note: *** in Sgnificance indicates robustness of the results, and for variables without a symbol, there is no evidence.

The robustness of the estimates was assessed through bootstrap resampling with 1,000 iterations, whose summary statistics are presented in **Table A5**.

Table A5. Bootstrap Statistics (1,000 Iterations).

Variable	Bootstrap Mean	Standard Deviation	2.5% CI	97.5% CI
Constant	555.44	11.74	532.62	579.02
SEI	18.84	1.27	16.46	21.38
Non-White Proportion	−67.97	12.82	−91.67	−41.40
Female Proportion	4.79	19.82	−31.85	46.87
Federal School	39.47	3.93	32.33	47.78
Municipal School	6.30	11.25	−16.02	32.30

To facilitate visual interpretation, **Figure A6** displays the empirical distributions of the coefficients estimated in each resampling. It can be observed that, for variables that are statistically significant in the model, the distributions are concentrated around the same sign as the original coefficient, indicating convergence among the estimates and confirming the consistency of the results.

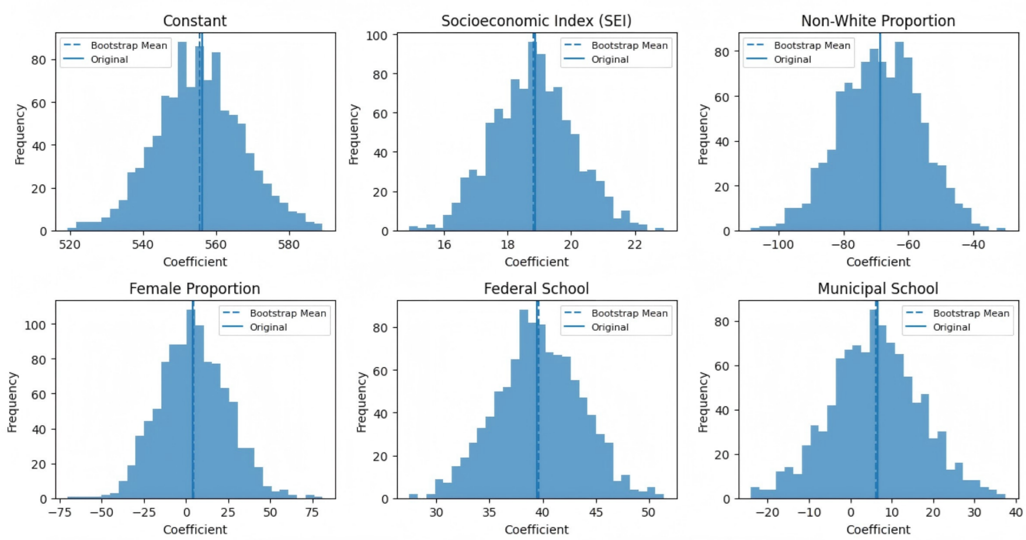

Figure A6. Bootstrap distributions of estimated coefficients.

This evidence points to the conclusion that the inferences presented in the main body of the work are statistically robust, maintaining validity even under specification adjustments and exclusion of influential observations.

Appendix D

The cluster analysis, based on socioeconomic, demographic, and institutional variables, identified four profiles of public schools in the state of Rio de Janeiro. **Figure A7** displays the centroids, allowing for a comparison of the average characteristics across the groups.

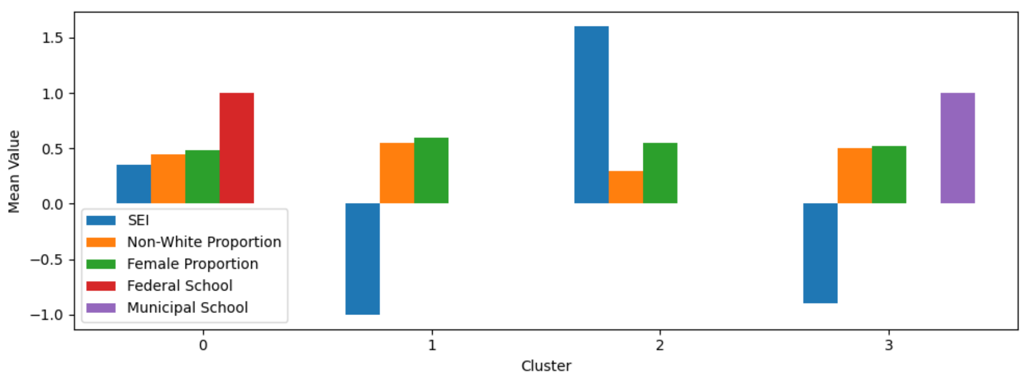

Figure A7. Visualization of cluster centroids.

The stability of the clusters was assessed using 100 bootstrap resamplings, with the Adjusted Rand Index (ARI) being calculated. **Table A6** presents the resulting statistics.

Table A6. Log-Linear Model Coefficients.

Metric	Value
Mean ARI	0.966
Standard Deviation	0.038
Q1 (25%)	0.942
Q3 (75%)	1.000
Minimum	0.868
Classification	High Stability

References

1. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Microdata from the 2023 ENEM Exam*; INEP: Brasília, Brazil, 2024. (in Portuguese)
2. Soares, J.F.; Collares, A.C.M. Family resources and cognitive performance by primary school students in Brazil. *Dados* **2006**, *49*, 615–650.
3. Coleman, J.S.; Campbell, E.Q.; Hobson, C.J.; et al. *Equality of Educational Opportunity*; U.S. Government Printing Office: Washington, DC, USA, 1966.
4. Becker, G.S. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*; National Bureau of Economic Research: New York, NY, USA, 1964.
5. Bourdieu, P. The forms of capital. In *Handbook of Theory and Research for the Sociology of Education*; Richardson, J., Ed.; Greenwood Press: New York, NY, USA, 1986; pp. 241–258.
6. Soares, J.F.; Alves, M.T.G. Racial inequalities in the Brazilian basic education system. *Educ. Pesqui.* **2003**, *29*, 147–165.
7. Franco, C.; Ortigão, I.; Alves, F.; et al. Quality and equity in education: Reconsidering the meaning of intra-school factors. *Rev. Bras. Educ.* **2007**, *12*, 5–26. (in Portuguese)
8. Melo, A.M.; Soares, J.F.; Xavier, F.P. Factors associated with ENEM performance: A multilevel analysis. *Rev. Bras. Estud. Popul.* **2021**, *38*, e0199.
9. Jolliffe, I.T.; Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150202. [CrossRef]
10. Baker, R.S.; Inventado, P. Educational data mining and learning analytics. In *Learning Analytics*; Springer: New York, NY, USA, 2014.
11. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *WIREs Data Min. Knowl. Discov.* **2020**, *10*.
12. OECD. *PISA 2022 Results*; OECD Publishing: Paris, France, 2023.
13. Bartlett, M.S. A note on the multiplying factors for various χ^2 approximations. *J. R. Stat. Soc. Ser. B Methodol.* **1954**, *16*, 296–298.
14. MacKinnon, J.G.; White, H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J. Econom.* **1985**, *29*, 305–325.
15. Jarque, C.M.; Bera, A.K. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econ. Lett.* **1980**, *6*, 255–259.
16. Breusch, T.S.; Pagan, A.R. A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **1979**, *47*, 1287–1294.
17. Cook, R.D. Detection of influential observation in linear regression. *Technometrics* **1977**, *19*, 15–18.
18. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June–18 July 1965, 27 December 1965–7 January 1966; pp. 281–297.
19. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.
20. Calinski, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **1974**, *3*, 1–27.
21. Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218.
22. Shapiro, S.S.; Wilk, M.B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52*, 591–611.
23. Ben-Hur, A.; Elisseeff, A.; Guyon, I. A stability-based method for discovering structure in clustered data. In Proceedings of the Pacific Symposium on Biocomputing, Hawaii, HI, USA, 3–7 January 2002; pp. 6–17.
24. IBGE. Educational indicators advance in 2024, but school delay increases. Available online: https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/43699-indicadores-educacionais-avancam-em-2024-mas-atraso-escolar-aumenta (accessed on 3 February 2026). (in Portuguese)
25. INEP. Brazil reached the highest percentage of full-time students in high school in 2025. Available online: https://www.gov.br/inep/pt-br/centrais-de-conteudo/noticias/censo-escolar/brasil-atingiu-maior-percentual-de-estudantes-em-tempo-integral (accessed on 3 February 2026). (in Portuguese)
26. Barbosa, J.A.F.; Cordeiro, R.L.F. Exploratory Analysis of Microdata from the National High School Exam-Enem: Performance and Specificities of the Participants. *J. Inf. Data Manag.* **2025**, *16*.
27. Fitzgerald, A.; Avirmed, T.; Battulga, N. Exploring the factors informing educational inequality in higher education: A systematic literature review. *Perspect. Policy Pract. High. Educ.* **2025**, *29*, 199–209. [CrossRef]

28. Moraes, C.P.; Peres, R.T. Reflections on performance diferences at ENEM: A socio-economic and schoolar analysis of Southeast Brazil. *J. Pol. Educ-s* **2022**, *16*, e85377.
29. Girão, J.E.d.O. Impacts of the pandemic and socioeconomic inequalities on the ENEM exam in Ceará (2019–2023). *UFC Inst. Repos*. **2023**.