ENEM under a Socioeconomic Perspective: Analysis and Evaluation Through Dimensionality Reduction

Cristiano C. Mendieta[1], **André L. Vignatti**[1]

[1]Departamento de Informática - Universidade Federal do Paraná (UFPR)
Curitiba – PR – Brasil

cristianomendieta@gmail.com, vignatti@inf.ufpr.br

Abstract. *This study investigates the relationship between socioeconomic factors and student academic performance in the 2022 ENEM, applying dimensionality reduction techniques to the microdata set provided by INEP. This dataset includes information collected from the exam, such as test scores, answer keys, evaluated items, participant scores, and responses to the socioeconomic questionnaire. The research compares linear methods, such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and Independent Component Analysis (ICA), with non-linear methods, such as Autoencoders and Pairwise Controlled Manifold Approximation Projection (PaCMAP), in binary and multiclass classification scenarios. The results indicate that linear methods provide a good balance between accuracy and computational efficiency, especially in binary classification scenarios. However, non-linear methods are more suitable for capturing complex structures in multiclass classifications, despite their higher computational cost. The Feature Selection technique using XGBoost proved effective in identifying key variables that differentiate students based on socioeconomic characteristics and academic performance. This study provides a comprehensive analysis of large educational datasets, generating results that can guide the formulation of public policies aimed at promoting equity within the Brazilian educational system.*

1. Introduction

The National High School Exam (ENEM), coordinated by the National Institute for Educational Studies and Research Anísio Teixeira (INEP), generates a vast set of microdata that integrates academic performance with detailed socioeconomic and demographic variables. While essential for analyzing factors associated with learning opportunities, the volume and heterogeneity of these data impose challenges such as informational redundancy, class imbalance, and computational inefficiency in machine learning models. In this scenario, dimensionality reduction is a useful strategy to simplify data representation and reduce computational cost, provided that its effects are interpreted in relation to the preprocessing and modeling decisions adopted in the study.

Dimensionality reduction is widely applied to mitigate data complexity across various domains, such as in Natural Language Processing with embeddings (*Word2Vec*, BERT) [Mikolov et al. 2013, Devlin et al. 2019], in biology with t-SNE and UMAP for cellular analysis [Becht et al. 2019], and in image processing via PCA and *Autoencoders* [Hinton and Salakhutdinov 2006]. In educational data mining, these techniques are particularly useful when large-scale assessments combine performance scores, demographic

descriptors, and socioeconomic questionnaires. In this work, we therefore use ENEM 2022 microdata to compare how different dimensionality reduction strategies preserve predictive information while reducing the complexity of the representation used by the classifier.

This work analyzes the application of dimensionality reduction techniques to ENEM 2022 microdata to evaluate the influence of socioeconomic factors on school performance. We investigated linear methods (PCA, SVD, ICA) and non-linear approaches (Autoencoders, PaCMAP) across binary and multi-class classification scenarios, using XGBoost for feature selection. To ensure clarity, our objective is twofold: (i) to compare dimensionality reduction strategies regarding their balance between data complexity and processing efficiency; and (ii) to identify the socioeconomic variables most consistently associated with performance-related outcomes. The results indicate that while linear methods significantly reduce training time without sacrificing accuracy, non-linear methods require more careful justification because their higher representational capacity does not necessarily translate into better performance for structured ENEM questionnaire data. This study thus supports equity-oriented analyses of educational disparities in Brazil.

The remainder of this article is organized as follows. Section 2 discusses related work; Section 3 presents the methodology; Section 4 reports the results; and Section 5 summarizes implications, limitations, and future work.

2. Related Work

In the educational field, the use of large volumes of data for analysis and modeling has expanded significantly, especially in large-scale assessments such as ENEM. The microdata provided by INEP encompass detailed information on academic performance and socioeconomic profiles, allowing for in-depth studies of factors associated with school success [Oliveira et al. 2024, Santos et al. 2023]. Recent studies on Brazilian educational inequalities have used data mining and dimensionality reduction to investigate the impact of socioeconomic factors on student performance, including the effects of the pandemic [Oliveira et al. 2024, Santos et al. 2023]. Queiroga et al. (2024) also reinforce that data-driven analyses can support school equity and inclusive public policies [Marques Queiroga et al. 2024].

From a methodological perspective, high-dimensional datasets impose severe challenges on the use of machine learning algorithms [Jia et al. 2022, Binois and Wycoff 2022]. The processing of these datasets faces the phenomenon known as the "curse of dimensionality," in which, according to Bellman (1957) [**?**], an excessive increase in dimensions makes the distances between points more uniform, hindering pattern identification and increasing computational costs. To mitigate redundancy and improve interpretability, dimensionality reduction techniques are commonly divided into two main approaches [Weikuan et al. 2022]: Feature Extraction, which transforms the original space into new variables or latent components that capture data variance and structure; and Feature Selection, which identifies and maintains only the most relevant subset of original variables.

In this work, feature extraction was applied using linear and non-linear methods. Linear methods assume that data can be projected into lower-dimensional subspaces via

linear combinations. Notable methods include Principal Component Analysis (PCA), which generates orthogonal components that maximize variance [Jolliffe 2002]; Singular Value Decomposition (SVD), which decomposes the original matrix into orthogonal components and singular values [Klema and Laub 1980]; and Independent Component Analysis (ICA), which seeks statistically independent components [Hyvärinen and Oja 2000]. On the other hand, non-linear methods capture complex relationships that linear transformations ignore. Pairwise Controlled Manifold Approximation Projection (PaCMAP) preserves local and global structures by adjusting the proximity between pairs of points [Wang et al. 2021], while Autoencoders utilize neural networks to compress data into a latent space and reconstruct them with minimal loss [Wang et al. 2012]. Feature selection in this study was performed via the XGBoost algorithm, which quantifies attribute importance through Gain, Cover, and Frequency (Weight) [Chen and Guestrin 2016].

Compared with studies focused on one family of techniques, our contribution is a controlled comparison of linear extraction, non-linear extraction, and model-based feature selection under the same classifier, targets, preprocessing pipeline, and component counts, enabling analysis of performance, cost, and interpretability.

3. Methodology

3.1. Dataset and Preprocessing

Large-scale data analysis requires specific care in data preparation to ensure the quality and reliability of the results. In this study, we worked with the microdata from the 2022 National High School Exam (ENEM), made available by the National Institute for Educational Studies and Research Anísio Teixeira (INEP). The preprocessing procedure involved multiple critical stages: the careful selection of relevant variables, handling of missing data, encoding of categorical attributes, and normalization of numerical values. These operations were essential to transform the raw set of 3,476,105 records and 76 attributes into a structured base suitable for the application of dimensionality reduction techniques and predictive modeling, while simultaneously maintaining the socioeconomic and academic performance information fundamental to the research.

Data Source and Characteristics: The ENEM dataset, obtained from INEP[1], includes a wide range of personal, social, economic, and performance information of the exam participants, covering variables such as age group, type of high school attended, parents' education level, presence of assets such as cars and refrigerators in the household, as well as the location of the completed school, among other characteristics. This dataset is suitable for comparing dimensionality reduction techniques. The original set contains 3,476,105 samples and 76 attributes and is provided in anonymized form, mitigating ethical concerns related to individual identification.

Variable Selection and Preparation: Variables were selected and prepared to create a consistent analytical dataset aligned with the study objectives. They were categorized as numerical (e.g., age, test scores, and year of high school completion) or categorical (e.g., sex, color/race, marital status, and socioeconomic questionnaire items). During preprocessing, variables were manually evaluated, and some columns were discarded because they did not add value to the problem under study or presented significant challenges

[1]https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

for efficient treatment. Columns containing the individual response vectors for each test were removed due to their complexity and the large number of possible combinations, for which effective encoding would be complex. Additionally, the column referring to the student's registration number was discarded as it provided no relevant information for the study's objectives, while the column with the school's municipality code was eliminated due to a high volume of null values. After this stage, the resulting dataset consisted of 54 columns (53 features plus the target variable). For transparency and reproducibility, details on attribute retention and removal are provided in the repository (see Sec. 3.3).

Handling of Missing Data: Missing data were handled by removing incomplete records. We adopted complete-case analysis to avoid imputation assumptions in a large and heterogeneous dataset; however, this choice may reduce representativeness and should be compared with imputation strategies in future work.

Encoding and Normalization: Categorical variables were encoded using the `OrdinalEncoder` from the *Scikit-learn* library, transforming them into numerical values. This keeps the representation compact for comparing several reduction methods, although ordinal encoding may impose artificial ordering on nominal categories. For numerical variables, normalization was applied via `MinMaxScaler`, scaling the values to the interval [0, 1]. This process is crucial to ensure that all variables have comparable weights in subsequent analyses.

Definition of Target Variables: Two classification scenarios were established to evaluate the effectiveness of dimensionality reduction techniques: *binary classification*, based on school type (public or private) and encoded as 0 for public school and 1 for private school; and *multi-class classification*, based on the family income bracket categorized into five levels (A, B, C, D, E), where A represents the highest income bracket and E the lowest. These targets represent complementary views of educational inequality: institutional segmentation and socioeconomic stratification. The multi-class classification was established according to the mapping of monthly family income into multiples of minimum wages, based on brackets where Class A covers incomes from 15 to 25 minimum wages, Class B from 10 to 15, Class C from 5 to 10, Class D from 2 to 5, and Class E up to 2 minimum wages, with all values derived from the minimum wage in effect in 2022. Although the Brazilian Institute of Geography and Statistics (IBGE) does not officially use the "Classes A to E" nomenclature, this categorization is widely applied in socioeconomic analyses and facilitates interpretation. Table 1 presents the target variables.

Table 1. Target Variable Definitions

Scenario	**Variable**	**Encoding**
Binary Classification	School Type	0: Public, 1: Private
Multi-class Classification	Family Income Bracket	A: Highest B: High C: Medium D: Low E: Lowest

3.2. Dimensionality Reduction Techniques

In this study, six dimensionality reduction techniques were implemented and compared, covering linear and non-linear methods, as well as the Feature Selection method. Table 2

provides a summary of these techniques and their primary parameters. The selected methods represent complementary assumptions: PCA, SVD, and ICA test linear combinations; Autoencoders and PaCMAP test non-linear transformations; and XGBoost-based feature selection tests whether retaining original variables improves interpretability without large predictive losses. For each technique, we tested different reduced dimensions: 2, 3, 5, 10, and 20 components, allowing for a comparative analysis of performance at various levels of reduction. These values cover severe, intermediate, and less restrictive compression levels.

Table 2. Dimensionality Reduction Techniques and Parameters Used

Technique	Type	Main Parameters	Values Used
PCA	Linear	svd_solver	Default value: "auto"
SVD	Linear	algorithm	Default value: "randomized"
ICA	Linear	max_iter, tol	max_iter: 200 (default), tol: 0.0001 (default)
PaCMAP	Non-linear	n_neighbors, MN_ratio, FP_ratio	n_neighbors: 10, MN_ratio: 0.5, FP_ratio: 2.0
Autoencoder	Non-linear	learning_rate, epochs, batch_size	learning_rate: 0.001, epochs: 200, batch_size: 32

Regarding the feature selection technique, the *XGBoost* algorithm was utilized, which evaluates variable importance during model training. This importance calculation is based on three main metrics: *Gain*, *Cover*, and *Frequency*. In this work, the model was initially trained with all variables available in the dataset, and the variables were ranked based on *Gain*. Later, new training sessions were conducted using only the most relevant subsets of variables identified by this metric. This approach allowed for a significant reduction in data dimensionality without compromising model performance, while also offering an interpretable analysis of the most critical variables for the classification task.

3.3. Modeling, Evaluation, and Reproducibility

In this section, we present the main points of the problem modeling and how the studied solutions were evaluated to assist in the comparative study. In the experiments conducted, all dimensionality reduction methods were applied to reduce the data to 2, 3, 5, 10, and 20 components. Following this, all datasets were submitted to the same classification algorithm: *XGBoost*. Furthermore, *XGBoost* was used with all 53 features of the resulting dataset (after the selection and preparation stage) as a "baseline" model, allowing for a comparison of the reduced models' results against the totality of the available data.

Classification Algorithm: The XGBoost classifier was chosen as the base algorithm to evaluate the performance of the different dimensionality reduction techniques. XGBoost (*Extreme Gradient Boosting*) is a machine learning algorithm based on decision trees, widely recognized for its efficiency and precision in classification tasks. It operates through an ensemble of trees, iteratively adjusting new models to correct errors from previous ones, which makes it robust for complex and high-dimensional datasets. Its ability to handle heterogeneous data and the use of advanced regularization techniques contribute to optimized performance and the prevention of overfitting, making it a reliable choice for a wide range of classification problems [Chen and Guestrin 2016].

Data Preparation for Modeling: The dataset was split into training (80%) and testing (20%) sets, with a fixed random seed to ensure reproducibility.

Training and Validation Process: The training and validation process followed three main steps: *Cross-Validation*, using the k-Fold method with 5 folds to ensure model stability and avoid overfitting; *Model Training*, where XGBoost was configured with 400 estimators and the objective functions binary:logistic for binary classification or multi:softmax for the multi-class scenario; and the definition of *Hyperparameters*, which were kept constant across all experiments to ensure a fair and standardized comparison between the different dimensionality reduction techniques.

Evaluation Metrics: To evaluate model performance, Accuracy was used—representing the proportion of correct predictions over the total predictions—along with the F1-score, which consists of the harmonic mean between precision and recall, providing a balanced measure of model performance. In addition to performance metrics, the training time was also recorded for each combination of reduction technique and number of components, allowing for an analysis of computational efficiency. Training time was measured as the elapsed wall-clock time of the reduction and classifier training stages under the same execution environment. Because values correspond to experimental runs rather than repeated benchmark averages, small differences should be interpreted cautiously.

Implementation and Environment: The implementation was carried out in Python 3.11, using the Pandas (1.2.4) and NumPy (1.20.2) libraries for data manipulation and structuring, and Scikit-learn (0.24.2) for preprocessing tools, evaluation metrics, and the implementations of the PCA, SVD, and ICA methods. For the PaCMAP method, the PaCMAP (0.6.3) library was used. TensorFlow (2.5.0) was used for the Autoencoder, XGBoost (1.4.2) as the classification algorithm, and Matplotlib (3.4.2) for result visualization and graphical analysis. The experiments were executed on the same workstation for all methods; hardware details are reported in the public repository to support reproducibility of the timing analysis.

Reproducibility: To ensure the reproducibility of the experiments, the following measures were adopted: the use of fixed random seeds in all stages involving randomness – such as data splitting and model initialization – and the provision of the complete source code, including preprocessing, training, and evaluation scripts, in a public GitHub repository[2]. This methodological approach enables a controlled comparative analysis of the different dimensionality reduction techniques applied to the ENEM 2022 data, providing valuable results regarding the effectiveness and efficiency of each method in the context of educational classification.

4. Results

The results of applying the different dimensionality reduction techniques were evaluated based on three main metrics: accuracy, *F1-Score*, and training time. Table 3 later summarizes the qualitative trade-offs observed across methods. For each technique, reduction was performed for 2, 3, 5, 10, and 20 components, and the dimensionally reduced data were used to train classification models using the XGBoost algorithm. The analysis was conducted in both binary and multi-class classification scenarios, allowing for a com-

[2]https://anonymous.4open.science/r/DimensionalityReductionENEM-134F

prehensive comparison of the techniques' performance across different dimensionality reduction configurations and classification settings.

4.1. Multi-class Classification: Family Income Prediction

This section analyzes dimensionality reduction in multi-class classification scenario (family income), focusing on test accuracy, F1-Score, and training time metrics.

Test Accuracy: The comparison of accuracies in Figure 1 highlights the balance of PCA and SVD, which exhibit consistent linear performance across all tested scenarios. The Autoencoder shows an evolution proportional to the number of components, matching PCA as it minimizes information loss during dimensional reduction. Similarly, ICA validates its effectiveness by reaching accuracy levels close to the other methods in higher dimensions, confirming its ability to extract relevant components for the multi-class task. PaCMAP showed inferior performance compared to linear methods, indicating that the structure of the multi-class problem may not require the complexity of a non-linear technique. Conversely, Feature Selection demonstrated robustness by achieving competitive results, reaching accuracy close to the *baseline*. This highlights that the direct identification of the most relevant variables is sufficient to maintain high accuracy, simplifying the model without significant loss of information.

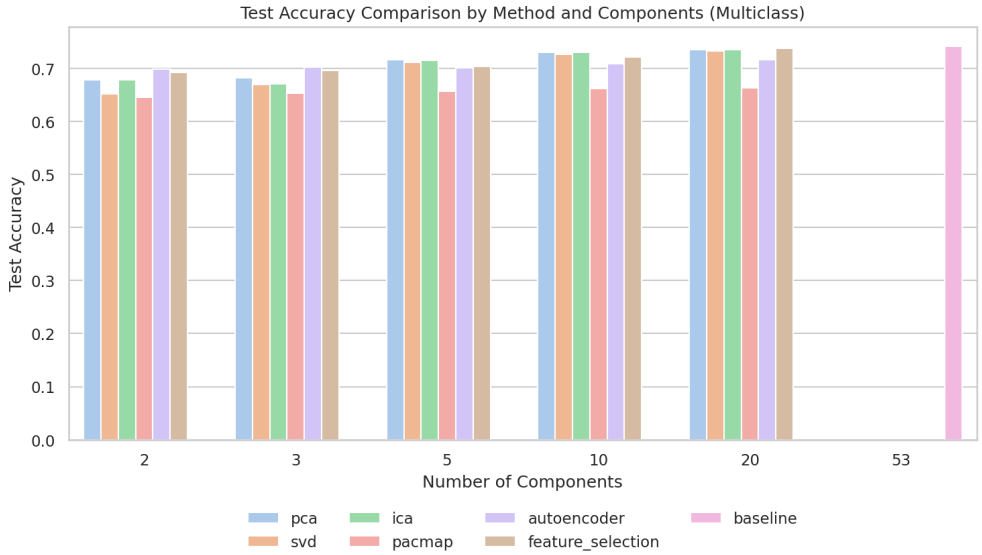

Figure 1. Accuracy by Method and Components in the Multi-class Scenario.

F1-Score: Figure 2 illustrates the stability of PCA, SVD, and ICA in the multi-class scenario, all presenting consistent and high *F1-Scores*. The *Autoencoder* follows this trend, demonstrating high information retention in higher dimensions. However, PaCMAP did not replicate the same success, indicating that the inherent complexity of its non-linear structure may not be necessary or effective for this problem. Notably, Feature Selection not only matched the performance of linear methods in low-reduction scenarios but also surpassed them in specific cases, validating the effectiveness of filtering irrelevant variables to optimize classification.

The stronger performance of linear methods suggests that the income-bracket signal is largely aligned with additive socioeconomic gradients, such as household assets,

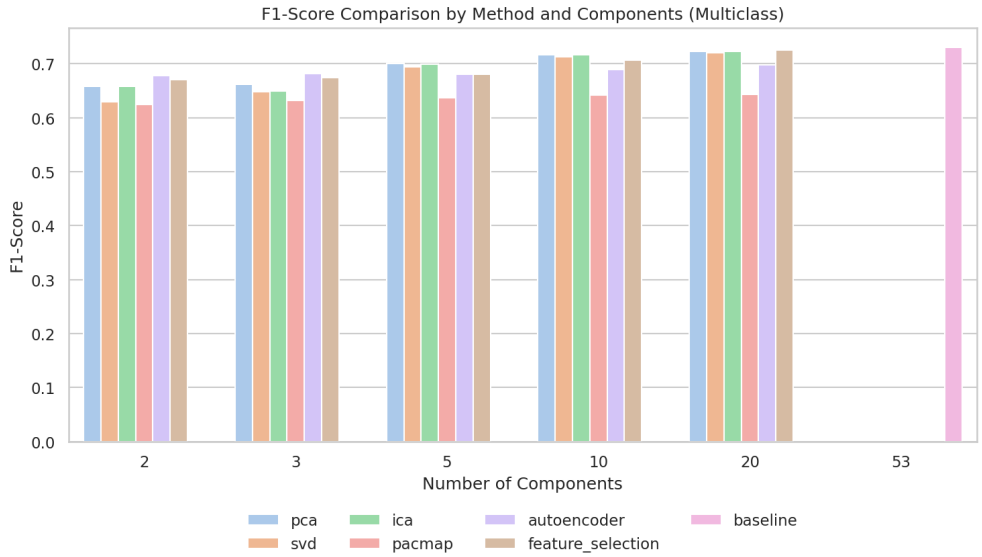

Figure 2. F1-Score by Method and Components in the Multi-class Scenario.

regional indicators, and educational background. The lower performance of PaCMAP indicates that neighborhood-preserving projections may be less suitable for supervised classification than for exploratory visualization.

Training Time: The analysis of Figure 3 reveals that PCA, SVD, and ICA maintain low and similar training times, followed closely by the *Autoencoder*, which demonstrates satisfactory computational efficiency. PaCMAP, in turn, does not offer competitive advantages in processing within this scenario. The outlier lies in Feature Selection, whose execution time is significantly higher, especially when attempting to drastically reduce the number of variables. Therefore, in systems where training agility is a critical requirement, linear methods (PCA, SVD, and ICA) stand out as the most balanced solutions.

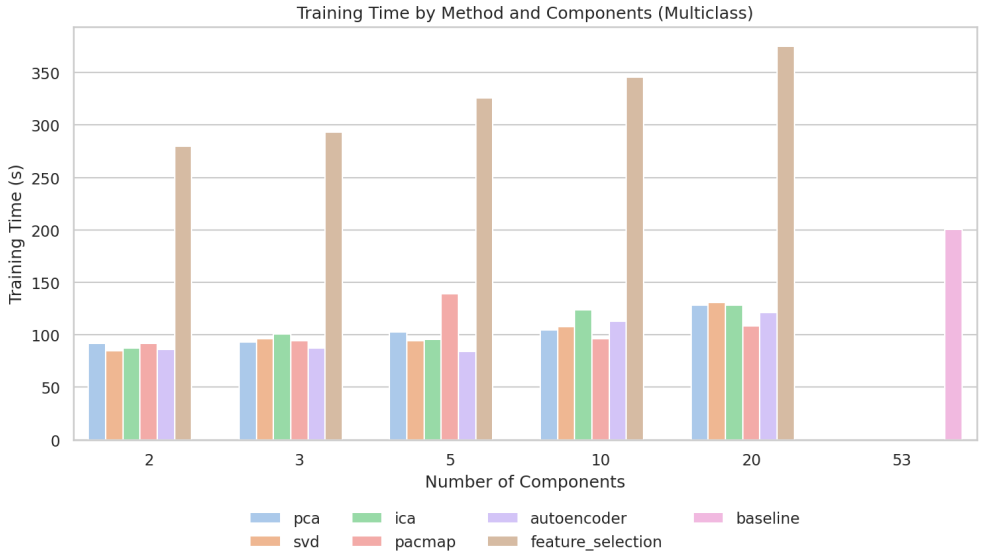

Figure 3. Training Time by Method and Components in the Multi-class Scenario.

4.2. Binary Classification: School Type Prediction

This section analyzes dimensionality reduction in binary classification scenario (school type), focusing on test accuracy, F1-Score, and training time metrics.

Test Accuracy: As shown in Figure 4, PCA and SVD prove to be solid choices for binary classification, with high and stable accuracies across different compression levels. The *Autoencoder* follows this trend, although it manifests slight instability when the number of components is drastically reduced. ICA also proves effective, capturing the essential variations for the task. On the other hand, PaCMAP reveals itself to be less suitable for this scenario of simplified relationships, presenting inferior performance. Finally, the Feature Selection technique reaffirms its effectiveness by offering competitive results, achieving outcomes close to the *baseline* and surpassing other methods in certain cases.

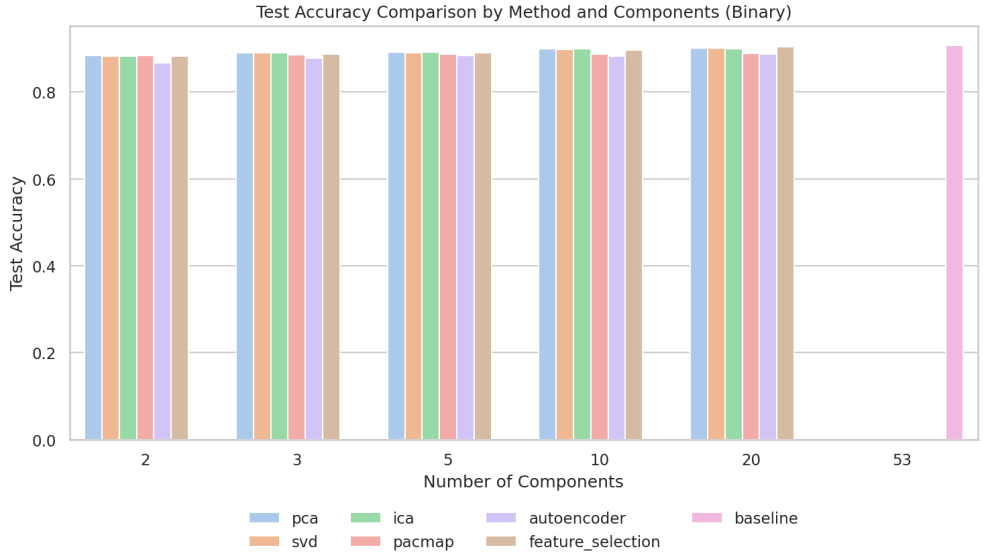

Figure 4. Accuracy by Method and Components in the Binary Scenario.

F1-Score: As seen in Figure 5, PCA, SVD, and ICA demonstrate high consistency in the binary scenario, reaching high *F1-Scores* with as few as five components. Interestingly, the *Autoencoder* and PaCMAP performed below expectations, indicating that the non-linear sophistication of these techniques may introduce noise or redundancy in simpler classification problems. Notably, Feature Selection achieved the best results in the series, surpassing the component extraction techniques and reaching values close to the *baseline*.

The binary task is comparatively easier because school type is strongly associated with fewer socioeconomic indicators. Thus, PCA, SVD, and ICA reach stable performance with few components, while the additional flexibility of non-linear approaches appears unnecessary for this target.

Training Time: Figure 6 shows the training time for the binary scenario. Linear methods, such as PCA, SVD, and ICA, continue to be the fastest. The *Autoencoder* presents training times comparable to linear methods, indicating acceptable computational efficiency. PaCMAP, especially with a larger number of components, requires significantly longer training time due to its computational complexity in preserving non-linear relationships. The high time observed with 20 components should be interpreted as a computational

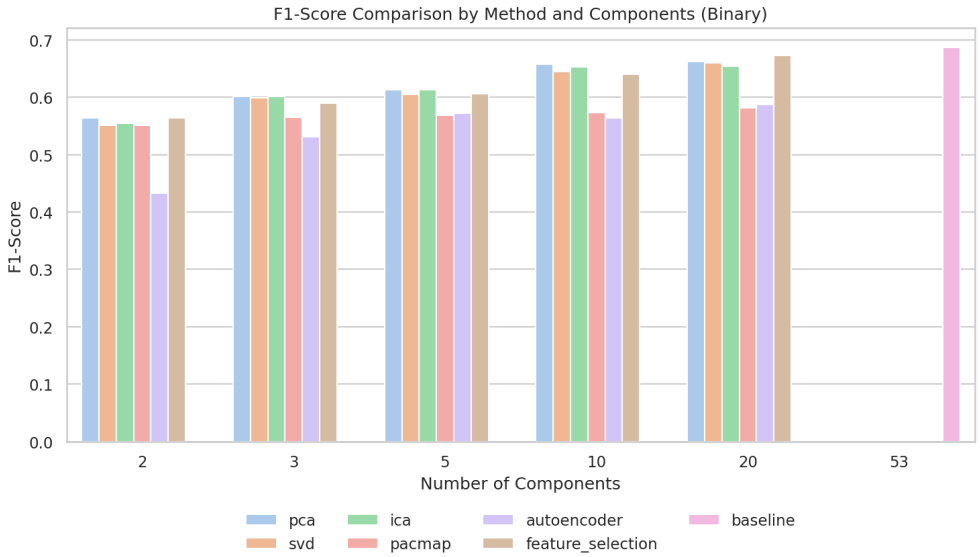

Figure 5. F1-Score by Method and Components in the Binary Scenario.

outlier of the manifold optimization step, not as evidence of additional value. Feature Selection presented a higher training time compared to the other methods, indicating that the selection process is computationally intensive. For applications where training time is critical, methods such as PCA, SVD, and ICA are more highly recommended.

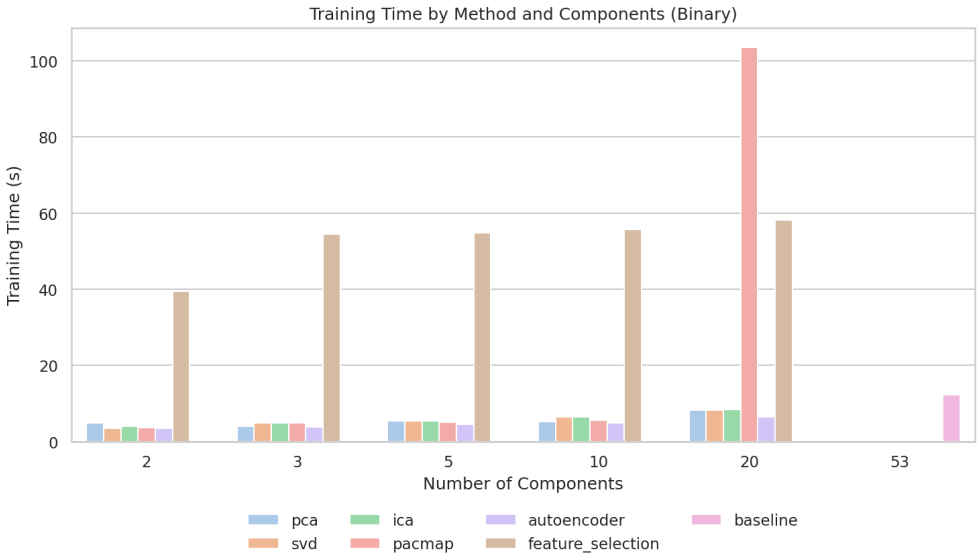

Figure 6. Training Time by Method and Components in the Binary Scenario.

4.3. Feature Selection Method Analysis

The Feature Selection technique was applied to identify the most influential variables for classification performance in both the binary and multi-class scenarios. Figures 7 and 8 present the most significant variables for each scenario, allowing for a detailed interpretation of the factors that most impact the model.

Binary Classification (School Type Prediction): For the binary classification scenario between public and private schools, the variables of highest relevance synthesize the contrast between academic performance and socioeconomic context. The `faixa_renda_familiar` (family income) consolidated as the most influential factor, reiterating that socioeconomic level is a critical determinant in access to private education, as corroborated by the literature [Sampaio and Guimarães 2009]. This profile is reinforced by indirect indicators of purchasing power and infrastructure, such as the presence of domestic workers (Q007), the availability of bathrooms (Q008), and access to computers (Q024). Regarding performance, scores in Writing Competencies 4 (NU_NOTA_COMP4) and 2 (NU_NOTA_COMP2) highlight disparities in argumentation skills, mastery of formal language, and textual comprehension, while the variable CO_UF_PROVA points to the impact of regional asymmetries. In summary, these findings reveal that the choice between public and private institutions in Brazil is driven by the convergence of the family's social stratum and the cultural and technological capital accumulated by the student, validating the strong correlation between material conditions and school trajectory.

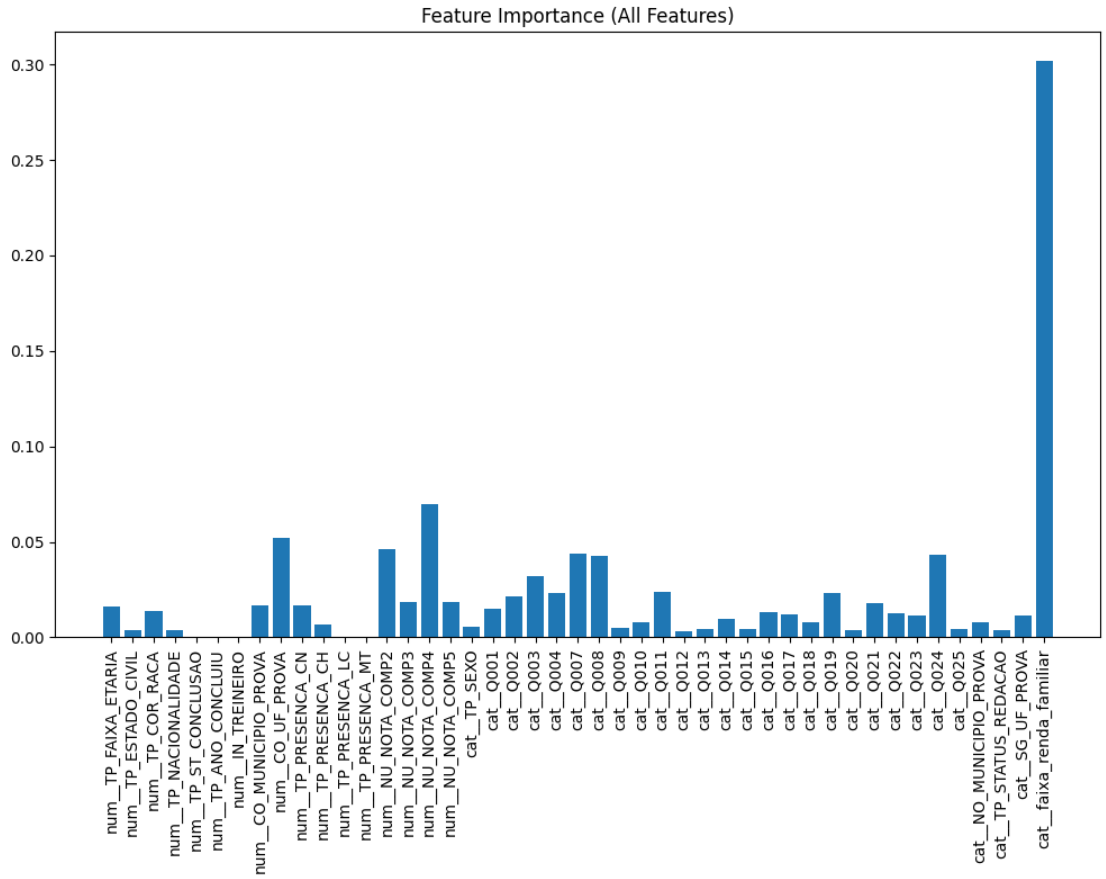

Figure 7. Feature Importance for the Binary Scenario.

Multi-class Classification (Family Income Bracket Prediction): For the multi-class classification scenario, focused on distinguishing family income brackets, the analysis reveals a predominance of socioeconomic and demographic variables over academic ones. The possession of durable consumer goods, such as a car (Q010), computer (Q024), washing

machine (Q014), and vacuum cleaner (Q018), consolidated as the primary set of indicators for purchasing power and domestic comfort, allowing for effective differentiation between middle and upper classes. Additionally, basic infrastructure indicators, such as the presence of a bathroom in the residence (Q008), and geographical factors, represented by CO_UF_PROVA, evidence how sanitation and regional economic disparities are determinants in Brazilian social stratification. These findings demonstrate that the Feature Selection method was effective in isolating critical variables that reflect the candidates' standard of living. Although it presents a higher computational cost, this technique proves advantageous when interpretability is essential, allowing for a deep understanding of the relationship between material conditions and the socioeconomic profile within the educational context.

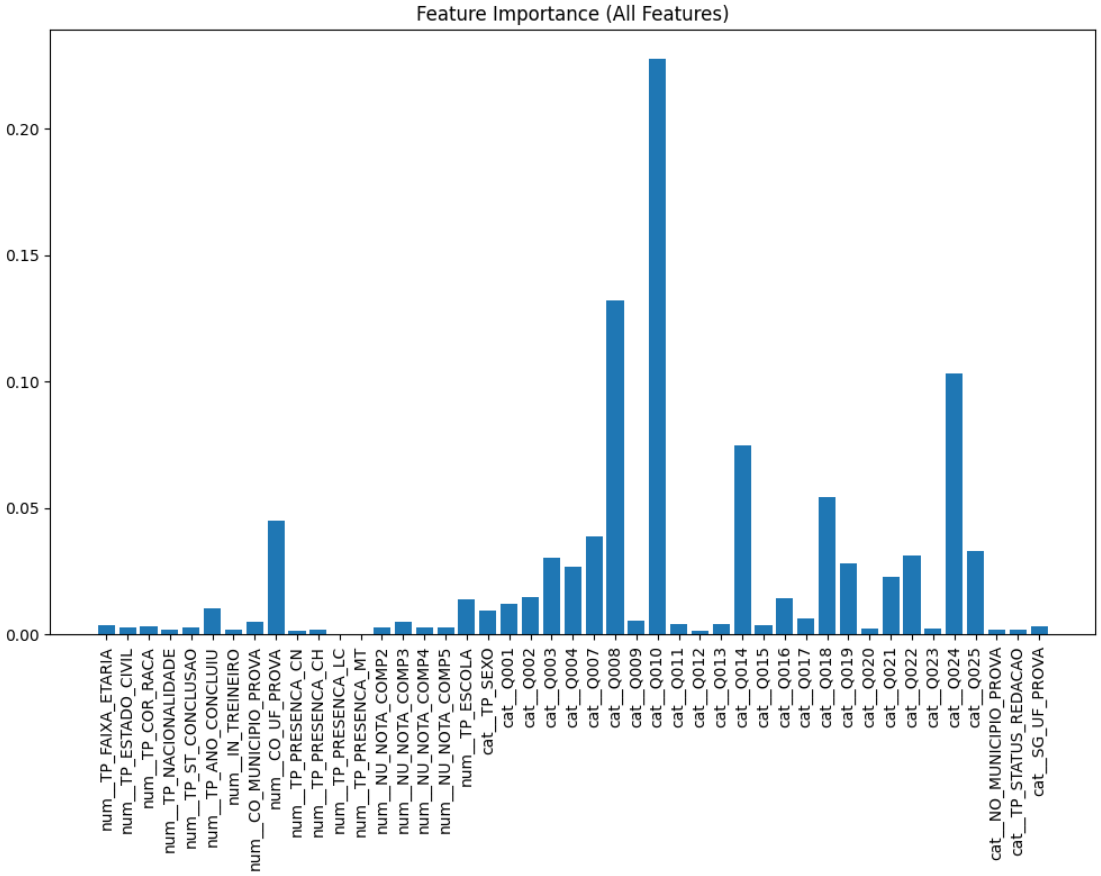

Figure 8. Feature Importance for the Multi-class Scenario.

5. Discussion and Final Considerations

Table 3 consolidates the main qualitative conclusions from the experimental results. Experiments conducted with the ENEM 2022 microdata demonstrate that linear methods (PCA, SVD, and ICA) offer the best balance between predictive metrics (*F1-Score* and accuracy) and computational efficiency. ICA, in particular, presented performance equivalent to PCA and SVD, evidencing that the assumption of independence between components was effective in capturing the variations pertinent to educational classification. In contrast, non-linear approaches exhibited distinct behaviors: the *Autoencoder* proved

competitive and robust in higher dimensions, while PaCMAP did not justify its high computational cost, suggesting that the structure of the analyzed data does not require the complexity of multi-scale non-linear mappings. Feature Selection consolidated as the technique with the highest precision, frequently surpassing extraction methods in both scenarios (binary and multi-class). Although it demands longer processing time, its ability to isolate critical variables – such as asset ownership and writing competencies – favors the interpretability of socioeconomic patterns. In summary, dimensionality reduction proved to be an effective strategy for simplifying large volumes of educational data, allowing for leaner models without significant loss of information.

Implications: The identified variables suggest that socioeconomic context remains strongly associated with educational outcomes. Regional variables (e.g., CO_UF_PROVA) should be interpreted with caution, as they may capture structural inequalities across states rather than isolated causal explanations.

Table 3. Qualitative summary of dimensionality reduction methods in this study.

Method	Predictive Performance	Training Time	Interpretability
PCA / SVD / ICA	High and stable	Low	Medium
Autoencoder	Competitive (higher dimensions)	Medium	Low
PaCMAP	Lower in our scenarios	High	Low
Feature Selection (XGBoost)	Highest in several cases	High	High

Limitations: The study is limited to ENEM 2022, which restricts temporal generalization. Complete-case analysis may change the final sample composition when missing values are not uniformly distributed, and ordinal encoding may impose artificial order on nominal variables. Finally, training times compare relative computational behavior under a common environment, but repeated benchmark averages would strengthen future timing analyses.

Final Remarks and Future Work: This study contributes a reproducible comparative pipeline for ENEM microdata analysis. Future work includes evaluating alternative missing-data strategies, comparing encodings for nominal variables, testing additional targets, replicating the pipeline in other ENEM editions, and analyzing geographic variables in greater detail.

References

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37(1):38–44.

Binois, M. and Wycoff, N. (2022). A survey on high-dimensional gaussian process modeling with application to bayesian optimization. *ACM Trans. Evol. Learn. Optim.*, 2(2).

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430.

Jia, W., Sun, M., Lian, J., and Hou, S. (2022). Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, 8(3):2663–2693.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, NY, 2nd edition.

Klema, V. and Laub, A. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, 25(2):164–176.

Marques Queiroga, E., Sarmanho Siqueira, E., Dos Santos Portela, C., Damasceno Cordeiro, T., Ibert Bittencourt, I., Isotani, S., Ferreira Mello, R., Muñoz, R., and Cechinel, C. (2024). Data-driven strategies for achieving school equity: Insights from brazil and policy recommendations. *IEEE Access*, 12:101646–101659.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Oliveira, E., Justo, W., and Lucena, M. (2024). The dynamics of public high school student performance in ceará: A study of the case of sobral in brazil. *IOSR Journal Of Humanities And Social Science*, 29:25–33.

Sampaio, B. and Guimarães, J. (2009). Diferenças de eficiência entre ensino público e privado no brasil. *Economia Aplicada*, 13(1):45–68.

Santos, B., Saporetti, C. M., and Macedo, B. S. (2023). Analysis of the impact of the pandemic on social inequalities in enem 2019 and 2020 using machine learning. *Semina: Exact and Technological Sciences*, 44(2):1–12.

Wang, J., He, H., and Prokhorov, D. V. (2012). A folded neural network autoencoder for dimensionality reduction. *Procedia Computer Science*, 13:120–127. Proceedings of the International Neural Network Society Winter Conference (INNS-WC2012).

Wang, Y., Huang, H., Rudin, C., and Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73.

Weikuan, Z., Qiang, L., and Jiawei, W. (2022). A comprehensive review on feature selection strategies for high-dimensional data. *Journal of Machine Learning Research*, 23(5):1–45.