

Context-Aware Learning Analytics for School Profiling Using SAEB Data

Leslie Miho Nobayashi¹, Glauco V. Pedrosa¹, Marcelo Ladeira¹

¹Graduate Program in Applied Computing (PPCA)
University of Brasilia (UnB) – Brasilia – DF – Brazil

leslie.miho@hotmail.com, {glauco.pedrosa, mladeira}@unb.br

Abstract. *This paper proposes a context-aware Learning Analytics framework for analyzing school performance using SAEB 2021 data. The proposed approach integrates PCA, IRT, XGBoost, and K-Means to characterize school profiles and identify contextual factors associated with Mathematics proficiency. The results revealed four groups of schools characterized by distinct determinants. Lower-performing schools were more strongly associated with student progression issues, whereas higher-performing schools were primarily associated with family support and infrastructure conditions. These findings reinforce the importance of context-sensitive educational policies and Learning Analytics for educational decision-making.*

Resumo. *Este artigo propõe um framework de Learning Analytics sensível ao contexto para análise do desempenho escolar utilizando dados do SAEB 2021. A abordagem integra PCA, IRT, XGBoost e K-Means para caracterizar perfis escolares e identificar fatores contextuais associados ao desempenho em Matemática. Os resultados identificaram quatro grupos de escolas com determinantes distintos. Escolas de menor desempenho apresentaram maior associação com problemas de progressão escolar, enquanto escolas de maior desempenho foram principalmente associadas ao apoio familiar e à infraestrutura. Esses achados reforçam a importância de políticas educacionais sensíveis ao contexto e do uso de Learning Analytics na tomada de decisão educacional.*

1. Introduction

Educational inequalities are influenced by contextual factors that affect schools differently. In heterogeneous educational systems such as Brazil's, schools with similar academic outcomes may present distinct contextual conditions and institutional needs [Silva et al. 2020, Sorgatto et al. 2021]. Because of this, policies that treat all schools the same often fail to meet the realities of different groups of schools [Ferreira et al. 2021].

In this context, Learning Analytics has emerged as an important field for supporting educational assessment and evidence-informed decision-making through educational data analysis. By combining data analysis, machine learning, and contextual interpretation, Learning Analytics enables the identification of patterns related to academic performance and educational inequalities [Romero and Ventura 2010, Kalita et al. 2025].

In the Brazilian educational context, the *Sistema de Avaliação da Educação Básica* (SAEB), conducted by the *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira* (INEP), represents an important source of educational data. In addition to cognitive assessments, SAEB provides contextual questionnaires that enable multidimensional analyses of school performance and educational conditions. These datasets have been increasingly explored in educational analytics studies investigating academic achievement and educational inequality [Farias et al. 2023, Santos et al. 2021].

Various analytical techniques have been employed in educational data analysis. Among these, Extreme Gradient Boosting (XGBoost) has demonstrated strong capability in identifying factors associated with academic performance [Wang and Luo 2024, Guevara-Reyes et al. 2025], while K-Means clustering facilitates the identification of groups of schools with similar characteristics [Liu 2022]. Furthermore, Item Response Theory (IRT) enables the estimation of latent constructs from educational questionnaire data, generating more reliable and interpretable indicators for the analysis of school contexts [de Ayala 2009, Robitzsch 2022].

Although previous studies have explored these approaches independently, there is still a lack of integrated Learning Analytics frameworks capable of combining contextual modeling, factor relevance estimation, and school profiling within a unified analytical pipeline [Sousa et al. 2021].

To address this gap, this paper proposes a context-aware Learning Analytics framework that integrates Principal Component Analysis (PCA), Item Response Theory (IRT), XGBoost, and K-Means clustering to identify latent contextual dimensions, determine factors associated with Mathematics proficiency, and characterize school profiles. The framework was evaluated using data from the 2021 Brazilian National Assessment of Basic Education (SAEB), focusing on public schools in the ninth grade of elementary education in the Federal District.

The proposed framework combines dimensionality reduction, psychometric modeling, machine learning, and school profiling within a unified analytical pipeline. This integration supports context-sensitive educational assessment and evidence-informed decision-making in large-scale educational systems. Furthermore, it enables the simultaneous analysis of contextual factors, their association with Mathematics proficiency, and the heterogeneity among schools, providing valuable insights for the design of targeted educational interventions.

2. Proposed Framework

This paper proposes a context-aware Learning Analytics framework for analyzing school performance through the integration of dimensionality reduction techniques, psychometric modeling, supervised machine learning, and unsupervised clustering. The proposed framework enables the identification of contextual factors associated with academic achievement while characterizing homogeneous school profiles based on educational context and student proficiency.

The proposed framework comprises three stages: latent contextual dimension construction, factor importance analysis, and school profiling. PCA and IRT are used to derive latent contextual indicators, XGBoost is employed to assess their importance

in explaining Mathematics proficiency, and K-Means clustering is applied to identify homogeneous school profiles. Together, these stages enable the analysis of contextual factors, their association with academic performance, and school heterogeneity, supporting evidence-based educational decision-making. Figure 1 illustrates the proposed framework.

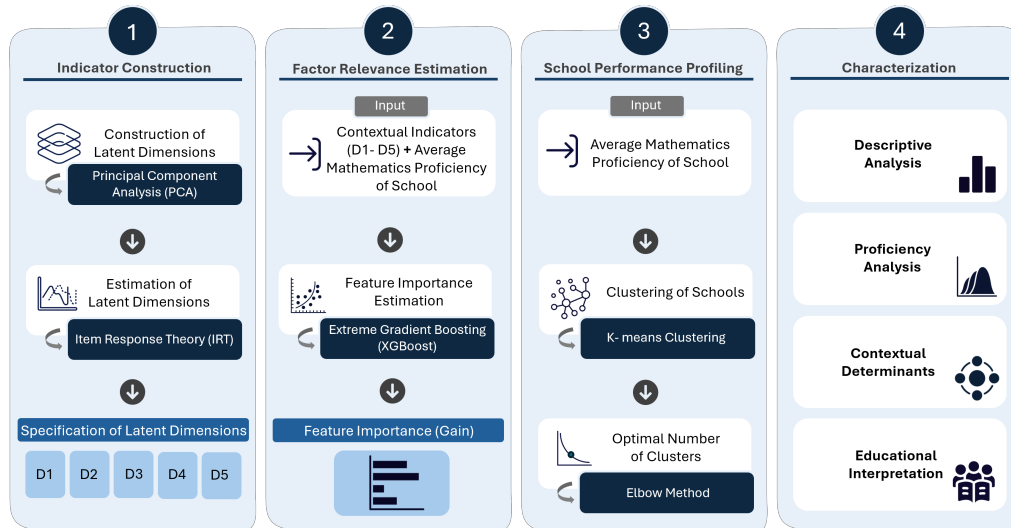


Figure 1. Proposed context-aware Learning Analytics framework.

2.1. Contextual Indicator Construction

The first stage of the framework consists of constructing continuous contextual indicators from questionnaire responses. Initially, Principal Component Analysis (PCA) was applied to identify the underlying structure of the data and define contextual dimensions based on the grouping of items with similar characteristics and the analysis of their factor loadings.

Next, considering that constructs such as family support, student progression, and adaptation during the COVID-19 pandemic are not directly observable, Item Response Theory (IRT) was employed to estimate the latent scores associated with each dimension. This approach provides more reliable and interpretable measures of the contextual factors under investigation.

In this study, the Graded Response Model (GRM), proposed by [Samejima 1969], was adopted due to its suitability for ordered categorical variables. The estimated latent scores were subsequently aggregated at the school level using arithmetic means, generating contextual indicators representative of each school.

2.2. Contextual Factor Relevance Estimation

After constructing the contextual indicators, the second stage of the framework identifies the contextual dimensions most associated with Mathematics proficiency. For this purpose, the XGBoost algorithm [Chen and Guestrin 2016] was adopted due to its robustness and ability to assess feature importance. The model was configured as a regression task, using the contextual indicators as predictor variables and the average Mathematics proficiency of each school as the target variable.

Feature importance was assessed using the gain metric extracted from the trained XGBoost models, enabling the ranking of contextual dimensions according to their contribution to Mathematics proficiency. The results provide interpretable evidence on the contextual factors most strongly associated with school performance and may support educational decision-making.

2.3. School Performance Profiling

The third stage of the framework consists of profiling schools according to their Mathematics proficiency levels using the K-Means algorithm [MacQueen 1967]. This stage aims to identify groups of schools with similar performance patterns, enabling the analysis of their contextual characteristics and the comparison of different educational profiles. Clustering was performed using the contextual indicators and the average Mathematics proficiency of each school, adopting Euclidean distance as the dissimilarity measure.

The optimal number of clusters was determined using the Elbow Method, which indicated four groups as the most appropriate solution. This configuration reduced within-cluster variability while maximizing differences among performance profiles. Finally, the consistency of the clusters was evaluated through Analysis of Variance (ANOVA), which revealed statistically significant differences among the mean proficiency levels of the clusters.

2.4. Cluster Characterization and Educational Interpretation

In the final stage, the clusters were analyzed to characterize school performance profiles and identify the contextual factors most relevant to each group. The analysis included descriptive statistics of contextual indicators and proficiency levels, comparisons of performance distributions across clusters, and interpretation of the relative importance of contextual factors associated with academic performance.

By integrating latent construct measurement, predictive modeling, and school profiling into a unified analytical framework, the proposed approach enhances the understanding of educational inequalities and supports the development of evidence-based educational policies and interventions.

3. Experimental Setup

3.1. Dataset

The experiments conducted in this study used data from the 2021 edition of the *Sistema de Avaliação da Educação Básica* (SAEB), provided by the National Institute for Educational Studies and Research Anísio Teixeira (INEP). The analysis focused on 9th-grade students from public schools in Brazil's Federal District.

SAEB combines standardized cognitive assessments with contextual questionnaires applied to students, teachers, principals, and schools. In this work, the average Mathematics proficiency score of each school was adopted as the target variable, while the contextual questionnaire responses were used to construct explanatory indicators associated with educational and socioeconomic conditions.

After preprocessing and consistency verification, the final dataset comprised 24,897 students distributed across 102 schools that satisfied the inclusion criteria. Schools

presenting incomplete contextual dimensions or invalid proficiency values were excluded from the analysis.

3.2. Contextual Dimensions

The contextual indicators used in the experiments were derived from students' questionnaire responses through Item Response Theory (IRT). Before the IRT estimation, Principal Component Analysis (PCA) was performed to examine the factorial structure of the questionnaire items and support the organization of latent contextual dimensions. Based on factor loadings, some items were reassigned or excluded due to low correlations, improving construct coherence and interpretability of the dimensions.

The resulting organization produced five latent contextual dimensions:

- **D1 – Parental Encouragement:** parental participation and support in students' academic life;
- **D2 – Neighborhood Infrastructure:** environmental and infrastructure conditions surrounding students' residences;
- **D3 – Socioeconomic and Infrastructure Characteristics:** socioeconomic and household infrastructure conditions;
- **D4 – School Progression Status:** indicators related to grade repetition and school dropout;
- **D5 – Student Adaptation During the Pandemic:** educational adaptation and access to resources during the COVID-19 pandemic.

For each dimension, student-level latent scores were estimated and subsequently aggregated at the school level through arithmetic averaging.

3.3. Preprocessing and Normalization

Before applying the clustering and predictive models, all contextual variables were standardized using Z-score normalization to ensure comparable scales across attributes:

$$z = \frac{x - \mu}{\sigma}$$

where x represents the original variable value, μ is the sample mean, and σ corresponds to the standard deviation.

Additionally, exploratory analysis identified two schools behaving as extreme outliers during clustering. Since these schools formed isolated singleton groups, they were removed from the clustering procedure to improve cluster stability and interpretability.

3.4. XGBoost Configuration

The XGBoost models were implemented using the `xgboost` package in the R programming language. The experiments adopted a regression setting in which contextual indicators were used as predictor variables and the average Mathematics proficiency of schools was used as the target variable.

The dataset was divided into training and testing subsets using an 80/20 split. Hyperparameter tuning was conducted empirically to improve predictive performance and avoid overfitting.

The final model configuration adopted the following parameters:

- Maximum tree depth: 2;
- Learning rate: 0.1;
- Subsample ratio: 0.75;
- Feature sampling ratio: 0.6;
- Minimum child weight: 1;
- Number of trees: 150.

Feature importance analysis was performed using the *gain* metric extracted from the trained XGBoost models.

3.5. Clustering Configuration

The optimal number of clusters was determined using the Elbow Method by evaluating the Within-Cluster Sum of Squares (WCSS) for different values of K . The analysis suggested that $K = 4$ produced an adequate balance between cluster compactness, separation, and interpretability. To reduce sensitivity to centroid initialization, the K-means++ strategy was adopted. The clustering algorithm was executed multiple times with different random seeds, and the partition with the lowest WCSS value was selected.

Table 1 summarizes the characteristics of the identified clusters. A one-way Analysis of Variance (ANOVA) was additionally performed to evaluate whether the identified clusters presented statistically significant differences regarding average Mathematics proficiency. The results rejected the null hypothesis ($p < 2 \times 10^{-16}$), indicating statistically distinct school groups.

Table 1. Characteristics of the Identified Clusters

Cluster	Proficiency Range	Number of Schools
Group 01	232–244	19
Group 02	245–253	32
Group 03	254–267	39
Group 04	268–279	12

3.6. Computational Environment

All experiments were conducted using the R programming language (version 4.3.0). The main libraries employed in the implementation were:

- `ltm` for Item Response Theory modeling;
- `xgboost` for predictive modeling and feature importance analysis;
- `cluster` for clustering analysis.

4. Results and Discussion

The feature-importance analysis revealed substantial differences in the contextual factors associated with Mathematics performance across the four school clusters (Figure 2). The results revealed that different contextual dimensions become more relevant depending on the proficiency level of each school group. This finding suggests that educational inequalities emerge from distinct contextual configurations and reinforces the need for differentiated educational policies tailored to specific school realities.

In the lowest-performing groups, school progression (D4), particularly factors related to grade repetition and dropout, emerged as the dominant dimension. Pandemic adaptation (D5) also showed relevant influence among vulnerable schools, suggesting persistent effects of educational disruption caused by COVID-19. In Group 02, parental encouragement (D1) became considerably more relevant, indicating that family support gains importance as severe structural barriers begin to be mitigated.

Cluster	Mathematics Proficiency	Most Influential Dimensions	Educational Implications
Low Group 01	232 - 244	D4 (<i>Student Progression</i>), D5 (<i>Pandemic Adaptation</i>)	Prioritize interventions aimed at reducing grade repetition and dropout rates, while promoting learning recovery.
Lower Intermediate Group 02	245 - 253	D4 (<i>Student Progression</i>), D1 (<i>Parental Support</i>)	Strengthen family engagement and support strategies to improve school attendance and retention.
Upper Intermediate Group 03	254 - 267	D5 (<i>Pandemic Adaptation</i>), D3 (<i>Socioeconomic Conditions</i>)	Reduce contextual inequalities and expand support resources available to students.
High Group 04	268 - 279	D2 (<i>Local Infrastructure</i>), D1 (<i>Parental Support</i>)	Maintain favorable family and community support conditions to sustain academic achievement.

Figure 2. Contextual determinants and hierarchy of educational needs across school performance profiles.

In contrast, the highest-performing groups exhibited a more multifactorial and context-dependent profile. Group 03 was strongly associated with pandemic adaptation (D5), socioeconomic conditions (D3), and surrounding infrastructure (D2), while Group 04 was primarily associated with environmental infrastructure and parental support.

Figure 2 summarizes the hierarchy of educational needs across school performance profiles. Lower-performing schools are mainly associated with student progression challenges, whereas higher-performing schools are more influenced by family support and infrastructure conditions. These findings reinforce the importance of context-sensitive educational policies tailored to different school realities.

From a policy perspective, these findings reinforce the importance of context-sensitive interventions. Educational strategies designed for high-performing schools may not be effective in vulnerable educational contexts where retention and progression issues remain unresolved. Similarly, policies focused exclusively on school retention may provide limited gains in contexts where such barriers have already been largely overcome.

5. Conclusion

This paper presented a context-aware Learning Analytics framework that integrates Principal Component Analysis (PCA), Item Response Theory (IRT), XGBoost, and K-Means clustering to analyze school performance using data from the 2021 Brazilian National Assessment of Basic Education (SAEB). The proposed approach enabled the identification of latent contextual dimensions, their association with Mathematics proficiency, and distinct school profiles based on proficiency levels and contextual characteristics.

The results revealed that the factors associated with school performance vary across proficiency levels. Lower-performing schools were more strongly associated with student progression challenges, particularly grade repetition and dropout, whereas higher-performing schools were associated with family support, infrastructure conditions, and access to educational resources.

By combining latent construct measurement, predictive modeling, and school profiling within a unified analytical framework, the proposed approach contributes to understanding educational heterogeneity and supports evidence-based educational decision-making.

This study has some limitations. The analysis was restricted to public schools in the Federal District and to data from the 2021 SAEB assessment, limiting the generalizability of the findings. In addition, the indicators were derived from self-reported contextual questionnaires and aggregated at the school level, which may introduce response biases and mask heterogeneity among students. Furthermore, K-Means assumes approximately spherical cluster structures and may be sensitive to centroid initialization.

Future research may extend the framework to other Brazilian states, educational levels, and SAEB assessment cycles, as well as investigate explainable Learning Analytics approaches capable of supporting more equitable and context-sensitive educational policies.

Declaration on the Use of Generative AI

The authors declare that generative Artificial Intelligence (AI) tools were used exclusively to support language revision, grammar improvement, and text refinement during the preparation of this manuscript. All scientific content, methodological decisions, analyses, interpretations, and conclusions were conceived, verified, and approved by the authors, who assume full responsibility for the content of this work.

References

- 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*, pages 785–794.
- de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. Guilford Press, New York.
- Farias, M. d. L., Gusmão, R. P. d., and Gusmão, C. S. D. (2023). Mineração de dados educacionais: Investigando a relação entre os microdados do inep e o desempenho do

- ideb. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1109–1119. SBC.
- Ferreira, L. A., Rodrigues, R. L., and Souza, R. N. P. M. d. (2021). Dados abertos educacionais brasileiros: Um mapeamento sistemático da literatura. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1186–1195. SBC.
- Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R., Cox-Riquetti, F., and Villegas-Ch, W. (2025). Machine learning models for academic performance prediction: interpretability and application in educational decision-making. *Frontiers in Education*, 10:1632315.
- Kalita, E., Oyelere, S. S., Gaftandzhieva, S., et al. (2025). Educational data mining: A 10-year review. *Discover Computing*, 28(1):81.
- Liu, R. (2022). Data analysis of educational evaluation using K-means clustering method. *Computational Intelligence and Neuroscience*, 2022:3762431.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297.
- Robitzsch, A. (2022). On the choice of the item response model for scaling PISA data: Model selection based on information criteria and quantifying model uncertainty. *Entropy*, 24(6):760.
- Romero, C. and Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 40(6):601–618.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4):100.
- Santos, V., Saraiva, M., and Oliveira, A. (2021). Educational data mining in brazil: A literature review on school dropout. In *Proceedings of the Brazilian Symposium on Computers in Education (SBIE)*.
- Silva, V. A. A. d., Moreno, L. L. O., Gonçalves, L. B., Soares, S. S. R. F., and Souza Júnior, R. R. (2020). Identificação de desigualdades sociais a partir do desempenho dos alunos do ensino médio no enem 2019 utilizando mineração de dados. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 72–81. SBC.
- Sorgatto, D. W., Nogueira, B. M., Cáceres, E. N., and Mongelli, H. (2021). Avaliação de classificadores para relacionar características escolares a indicadores educacionais. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1232–1242. SBC.
- Sousa, E. B. G. d., Alexandre, B., Ferreira Mello, R., Pontual Falcão, T., Vesin, B., and Gašević, D. (2021). Applications of learning analytics in high schools: A systematic literature review. *Frontiers in Artificial Intelligence*, Volume 4 - 2021.
- Wang, S. and Luo, B. (2024). Academic achievement prediction in higher education through interpretable modeling. *PLOS ONE*, 19(9):e0309838.