International Conference
The Future of Education

Prediction of Students Performance Based on School Dropout Risk Factors

Leogildo Alves Freires[1], Heitor Marinho da Silva Araújo[2], Julio Cezar Albuquerque da Costa[3], Luan Filipy Freire Torres[1], Gabriel Macedo[1], Ane Mayra Melo Silva[1]

[1]Federal University of Alagoas, Brazil
[2]Federal University of Bahia, Brazil
[3]Federal University of Minas Gerais, Brazil

Abstract

School dropout has been widely recognized as a complex and multifactorial phenomenon, with significant implications for both individual development and educational systems. In recent years, advances in early warning systems have sought to anticipate this risk by identifying factors associated with school disengagement, particularly those of a relational nature, such as interactions among students, families, and schools [1]. Study aimed to investigate the predictive effect of school dropout risk factors on the performance of Brazilian students, using a synthetic sample of 10,000 cases, derived from a real dataset of 3,678 students from four states of Brazil. Synthetic data was generated using the Gaussian Copula technique, which estimates new responses based on the correlation between the observed variables in the real dataset, reproducing its relationships. Due to ethical considerations and compliance with Brazilian data protection legislation (LGPD), this study relies on synthetic, anonymized data, however, a nationwide data collection using real-world data is currently being conducted in Brazil. This approach was considered due to the General Law of Data Protection (LGPD). Results show that the relationships of students with family, other students, and school professionals' impacts on school performance. Results indicated that the pedagogical quality (p=.004), parenting (p<.001), support structure at home (p<.001), coordination with the family (p=.048), and belonging (p<.001) impact students' performance. These findings reinforce the central role of relational dimensions in shaping students' academic outcomes, highlighting that school performance is not solely determined by individual or structural factors, but emerges from the quality of interactions within the educational ecosystem. Overall, results provide empirical support for the advancement of early warning systems that incorporate relational indicators, contributing to more sensitive, context-aware, and effective strategies to prevent school dropout.

Keywords: *School dropout, Large-scale analysis, Educational psychology*

1. Introduction

School dropout has been widely recognized as a complex and multifactorial phenomenon, with profound implications for both individual development and the structuring of educational systems. The advancement of early warning systems has sought to anticipate this risk by identifying factors associated with school disengagement, overcoming purely individualistic views to focus on relational dimensions [1]. Understanding these interpersonal and institutional dynamics is fundamental for the design of more assertive public policies and for the promotion of a protective educational ecosystem that addresses the root causes of performance inequalities.

In order to capture the complexity of these interactions in different intergroup contexts, the use of robust and ecologically valid psychometric instruments is essential. Assessment models have been developed and improved, structuring risk and protective factors in various dimensions of student life. The refinement of these measures, as observed in the adaptations of the Relational Factors Scale for the Risk of School Dropout in its revised and alternative versions [2],[3] allows for the multivariate mapping of specific vulnerabilities that permeate the student's relationship with school professionals, their family structure, the local community and their peers.

Despite the urgency, contemporary empirical research faces the challenge of balancing investigative rigor with ethical privacy guidelines; therefore, handling educational microdata from vulnerable populations requires innovative approaches. A methodological alternative to the generation of synthetic data has emerged as a promising methodological solution, using techniques such as Gaussian Copula [4],[5] to reproduce the relationships and distributions of the original data [6],

Commented [1]: Adicionar @heitor.araujo@nees.ufal.br _Assigned to heitor.araujo@nees.ufal.br_

preserving the statistical integrity necessary for inferential analyses without compromising the anonymity of the participants.

Advances in the formulation of early warning systems have challenged strictly individualistic explanatory models, highlighting that the risk of academic disengagement emerges from vulnerabilities intrinsic to the relational networks in which the individual is immersed, particularly in the interactions established with the school ecosystem and the family nucleus. The investigative gap that guides the problematization of this work therefore lies in the need to measure the predictive effect of these psychosocial risk factors on the concrete performance of students.

Given the above, an assessment of the level of international vulnerability was conducted using the Alternative Version of the Scale of Relational Factors for the Risk of School Dropout (IAFREE-A). The analytical approach focused on examining the extent to which these relational dimensions statistically differentiate students with normative trajectories from those who have already experienced grade repetition. Through non-parametric inferences for independent groups, our hypothesis is that the history of academic failure is linked to weaknesses in interactions, constituting an expression of broader socio-educational disinvestment.

2. Method

2.1 Participants

The participants were students enrolled in public schools in four Brazilian states, comprising an original dataset of 3,678 individuals, collected as part of the School Trajectory Protection System (SPTE), a Brazilian initiative aimed at protecting students' educational trajectories. From this original dataset, a synthetic sample of 10,000 cases was generated, in which females predominated (50.82%), students concentrated in the Midwest region (44.35%), and 9,536 (95.4%) were classified as non-repeaters and 458 (4.6%) as repeaters (Table 1).

Table 1.

Sex	n	%	Race/Ethnicity	n	%
Male	4898	48.98	White	2168	21.68
Female	5082	50.82	Black	2062	20.62
I prefer not to declare	20	0.20	Brown	5488	54.88
			Other	282	2.82
Region of Brazil	n	%	Failure history	n	%
Midwest	4435	44.35	Repeaters	458	4.60
Northeast	1368	13.68	Non-repeaters	9536	95.40
North	1625	16.25			
Southeast	2572	25.72			

Characteristics of the participants' sociodemographic background.

2.2 Instruments

The final version of the Relational Factors for the Risk of School Dropout Scale – Alternative Version (IAFREE-A) [3] is an adaptation of the instrument originally developed by [1] (IAFREE) and revised by [2] (IAFREE-R), structured based on the concept of School Trajectory Protection. The instrument, completed by students, consists of 46 items organized into five protective relational dimensions: Student–School (SSc), Student–School Professionals (SP), Student–Family (SF), Student–Community (SC), and Student–Student (SSt). These dimensions are distributed across 14 factors of attention: Infrastructure (SSc1), School as a Safe Space (SSc2), Discrimination (SSc3), School Management and Organization (SP1), Pedagogical Quality (SP2), Teachers' Expectations (SP3), Parenting (SF1), Family Support Structure (SF2), Family Interaction (SF3), Network Interaction (SC1), Community Relations (SC2), Interpersonal Relations and Social Skills (SSt1), Expectations Regarding Education/Educational Development (SSt2), and Belonging/Identification (SSt3). Responses are recorded on a four-point Likert scale, presented in two formats: frequency ("Never" to "Always") and agreement ("Strongly Disagree" to "Strongly Agree"), both accompanied by pictorial cues. For most items, the scoring is reversed, with the exception of Q5 and Q36, which are phrased negatively. Thus, higher scores indicate a greater need for attention to academic progress, while lower scores indicate a lower degree of vulnerability. To characterize the participants' profiles, a 33-item sociodemographic questionnaire was also administered, covering variables such as gender, race/ethnicity, and history of failing grades, among others.

2.3 Procedures

Data collection was conducted in operational cycles, which provide schools with two options for administering the IAFREE-A: a digital version, accessible via a platform integrated into tablets and computers, and a printed version, intended for institutions with limited or unstable internet connectivity. In this context, students from schools belonging to four state school systems located in the Central-West, Northeast, North, and Southeast regions of Brazil participated in the study. The research was conducted in accordance with ethical guidelines for studies involving human subjects, as per Resolutions No. 466/2012 and No. 510/2016 of the National Health Council, and was approved by the Human Research Ethics Committee (CEP/UFAL), under opinion No. 5.407.594. Furthermore, all procedures related to the collection, storage, and processing of information followed the principles established by the National Council of Ethics in Research (CONEP) and the provisions of the General Data Protection Law (LGPD) [7], ensuring confidentiality, anonymity, and protection of participants throughout the entire research process.

2.4 Data Analysis

The analyses were performed using the R programming language [8], employing the Gaussian copula technique to generate synthetic data. Based on copula theory, this method allows for the modeling of multivariate distributions by separating the marginal distributions from the dependency structure among the variables [4] [9], generating new synthetic observations from integral probability transformations applied to real data [5]. Among its main advantages is the preservation of associations between variables in the database, maintaining the functional equivalence of synthetic data for inferential purposes [10], a relevant aspect in psychometric studies, in which latent factors are derived from correlations between items [11]. Thus, the similarity between the real and synthetic data was assessed using the Hellinger distance, an indicator ranging from 0 to 1, where values close to 0 represent greater similarity between the distributions [6], with values up to 0.20 considered adequate.

Based on this, the aim was to examine whether, and to what extent, the relational factors of the IAFFREE-A differentiate students according to their academic performance trajectory, specifically with grade repetition. Thus, group differences were operationalized by forming two groups defined by academic performance history: students who had never failed a grade (non-repeaters) and students who had repeated at least one school year (repeaters), measured using a single dichotomous self-report item: "Have you ever failed a grade?", to which participants responded with (0) "No, I have never failed" or (1) "Yes, at least once." This binary classification served as the independent variable, while the scores on the dimensions and factors of the IAFREE-A constituted the dependent variables of interest.

The differences were examined using the Mann–Whitney U test (also known as the Wilcoxon rank-sum test; [12] [13], a nonparametric procedure suitable for comparing two independent groups when the distributional assumptions of parametric tests cannot be met. Thus, the magnitude of group

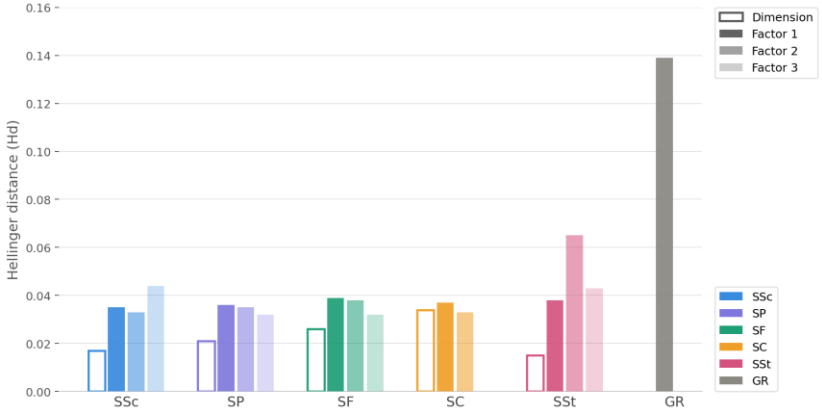
differences was estimated using the rank-based biserial correlation (r) as a measure of effect size for the Mann–Whitney U test [14]. Effect sizes were classified according to [15] as small (|r| ≥ 0.10), medium (|r| ≥ 0.30), and large (|r| ≥ 0.50). All comparative statistical analyses were conducted using JASP [16].

3. Results

The quality of the correspondence between the synthetic data and the original data was assessed using the Hellinger distance (Hd), a measure ranging from 0 to 1, where values closer to zero indicate greater similarity between the distributions analyzed. In this regard, all values obtained remained below the established adequacy criterion (Hd < 0.20), demonstrating high similarity between the distributions of the synthetic and original data.

As shown in Figure 1, the Hd values ranged from 0.015 (SSt) to 0.034 (SC). At the factor level, a slightly more significant range of variation was observed, with values between 0.032 (SP3 and SF3) and 0.065 (SSt2), all remaining within the limits considered satisfactory, which corroborates the quality of the approximation achieved. As for the sociodemographic variable, Grade repetition recorded Hd = 0.139, a value that, although higher than the others, remained below the adopted cutoff point, possibly due to the marked imbalance in the distribution of this variable in the original sample, characterized by a low proportion of students repeating a grade, a condition that tends to increase the complexity of its synthetic reproduction compared to variables with more balanced distributions.

Fig. 1. Hellinger distance by dimension, factor, and grade repetition.



Note. Hd: Hellinger distance. Colors indicate dimension: Darker shade = Factor 1; medium shade = Factor 2; lighter shade = Factor 3; outlined bar = Dimension. GR = Grade Repetition. SSc: Student–School; SP: Student–School Professionals; SF: Student–Family; SC: Student–Community; SSt: Student–Student.

The results of the Mann-Whitney test indicated statistically significant differences between students who had repeated a grade and those who had not across seven variables in the questionnaire (Table 2). Students with a history of grade repetition reported more negative perceptions regarding family support, family involvement, school belonging, and interpersonal relationships compared to those who had never repeated a grade. The most significant differences were concentrated particularly in dimensions related to family dynamics and identification with the school, indicating that grade repetition is not limited solely to low academic performance but is also associated with broader relational vulnerabilities that permeate students' school trajectories.

As for students who had to repeat a grade, the results showed higher means for the factors SF1 (M = 1.983; SD = 0.578), SF2 (M = 1.911; SD = 0.569), SF3 (M = 2.427; SD = 0.694), SSt2 (M = 1.718; SD = 0.492), in the SF dimension (M = 2.114; SD = 0.485), and in the SSt dimension (M = 2.099; SD = 0.525). The concentration of differences in the factors of the SF dimension indicates that the family context is the most sensitive domain to the pattern of academic failure, suggesting that students who repeat a grade perceive less support, less structure, and more distant family interactions regarding school life. The difference in SSt2, in turn, indicates that these students also have lower expectations

regarding their own educational development, which may reflect a process of progressive disengagement from their school trajectory and academic performance.

With regard to students who did not repeat a grade, higher means were observed only for factors SP2 (M = 2.307; SD = 0.584) and SF3 (M = 2.491; SD = 0.660). The result for SP2 suggests a more critical assessment of pedagogical quality by these students, given their greater engagement with the learning process; conversely, the difference in SF3 warrants interpretive caution, given the trivial effect size observed.

In this regard, effect sizes ranged from trivial to small, with small effects observed in SF1 (r = .125), SF2 (r = .113), and SSt2 (r = .154), while SP2 (r = −.070), SF3 (r = −.058), SF (r = .071), and SSt (r = .064) showed magnitudes below the cutoff point for a small effect. Among the variables analyzed, SSt2 showed the most significant effect size, although still of small magnitude, suggesting that students' expectations regarding their own educational trajectory constitute the relational factor with the greatest discriminative power between the groups.

Table 2. Mann–Whitney comparison tests: school dropout risk factors and the performance of Brazilian students.

Variables	Groups	Mean	SD	*U*	*p*	Effect Size
SP2 – Pedagogical quality	*1 = "No, I have never repeated"*	2.307	0.584	2.912	.009	−.070
	2 = "Yes, at least once"	2.226	0.556			
SF1 – Parenting	*1 = "No, I have never repeated"*	1.858	0.569	-4.611	<.001	.125
	2 = "Yes, at least once"	1.983	0.578			
SF2 – Home support structure	*1 = "No, I have never repeated"*	1.799	0.575	-4.067	<.001	.113
	2 = "Yes, at least once"	1.911	0.569			
SF3 – Engagement with the family	*1 = "No, I have never repeated"*	2.491	0.660	1.978	.034	−.058
	2 = "Yes, at least once"	2.427	0.694			
SSt2 – Belonging / Identification	*1 = "No, I have never repeated"*	1.589	0.487	-5.511	.001	.154
	2 = "Yes, at least once"	1.718	0.492			
SF - Student-Family	*1 = "No, I have never repeated"*	2.054	0.494	-2.539	.010	.071
	2 = "Yes, at least once"	2.114	0.485			
SSt - Student-Student	*1 = "No, I have never repeated"*	2.033	0.523	-2.637	.021	.064
	2 = "Yes, at least once"	2.099	0.525			

Note. SD = Standard Deviation. *U* = Mann–Whitney U test statistic. Effect Size = rank-biserial correlation. $p < .05$.

4. Discussion

As demonstrated, the family microsystem plays a central role in structuring trajectories of academic vulnerability [1]. The concentration of statistically significant differences in structural support factors at home and in parenting highlighted how grade repetition goes beyond strictly individual cognitive deficits, becoming the outcome of broader relational deprivations for students. The family, therefore, acts as the most sensitive domain in relation to the pattern of school failure. For the formulation of public policies, we indicate that the design of systems to protect school trajectories must

Commented [2]: Adicionar @heitor.araujo@nees.ufal.br _Assigned to heitor.araujo@nees.ufal.br_

transcend intra-classroom barriers, requiring integrated socio-educational strategies that foster the structuring and engagement of the family unit.

Regarding intergroup relations and institutionalization, belonging and identification with the school revealed discriminatory power among student profiles. Students with a history of grade repetition reported significantly lower expectations regarding their own educational development, as well as more negative perceptions about their belonging to the school environment. This pattern illustrates the crystallization of a progressive disengagement process, in which distancing from normative support networks reinforces the phenomenon of exclusion. The mitigation of exclusionary interpersonal dynamics and the strengthening of intergroup cohesion in the school climate thus emerge as crucial vectors to ensure that academic performance is supported by quality protective interactions.

Counterintuitively, we note that students with academic records free from failing grades reported more critical perceptions regarding the pedagogical quality offered by the institution. This dynamic suggests that greater engagement with the learning process refines the student's evaluative criteria, turning them into a more demanding observer of the infrastructure and teaching practices. Adherence to and success in the educational system presuppose an active demand for excellence and continuous support, challenging educational management to constantly refine its institutional approaches to meet the needs of highly engaged students.

REFERENCES

[1] Vasconcelos A. N., Freires L. A., Loureto G. D. L., Fortes G., Costa J. C. A., Torres L. F. F., Bittencourt I. I., Cordeiro T. D., Isotani S., "Advancing school dropout early warning systems: The IAFREE relational model for identifying at-risk students", Frontiers in Psychology, Vol. 14, 2023. https://doi.org/10.3389/fpsyg.2023.1189283
[2] Costa J. C. A., Torres L. F. F., Loureto G. D. L., Freires L. A., Freitas A. L. G. B., Júnior N. A. T., et al., "Relational Factors for the Risk of School Dropout Scale-Revised (IAFREE-R)", Iberoamerican Journal of Health and Social Research, 4, 2026, pp. 50-73.
[3] Freires L. A., Costa J. C. A., Torres L. F. F., Loureto G. D. L., Macêdo G. F. C., Silva A. M. M., Freitas A. L. G. B., Moro A., Cordeiro T. D., "Relational Factors for the Risk of School Dropout Scale - Alternative Version: Adaptation and evidence of validity", PLOS ONE, 2026, in press.
[4] Sklar A., "Fonctions de répartition à n dimensions et leurs marges", Publications de l'Institut de Statistique de l'Université de Paris, 8, 1959, pp. 229–231.
[5] Li Z., Zhao Y., Fu J., "Sync: A copula based framework for generating synthetic data from aggregated sources", 2020 International Conference on Data Mining Workshops (ICDMW), IEEE, 2020, pp. 571-578. http://doi.org/10.1109/ICDMW51313.2020.0008
[6] Mosquera L., El Emam K., Ding L., Sharma V., Zhang X. H., Kababji S. E., Eurich D. T., "A method for generating synthetic longitudinal health data", BMC Medical Research Methodology, Vol. 23, No. 1, 2023, p. 67.
[7] Brasil, "Lei nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais)", Diário Oficial da União, Brasília, 2018.
[8] R Core Team, "R: A Language and Environment for Statistical Computing", Vienna, R Foundation for Statistical Computing, 2025. Available at: https://www.R-project.org/
[9] Nelsen R. B., An Introduction to Copulas, 2nd ed., Springer, New York, 2006.
[10] Ahmadian M., Bodalal Z., van der Hulst H. J., Vens C., Karssemakers L. H., Bogveradze N., Castelijns J. A., "Overcoming data scarcity in radiomics/radiogenomics using synthetic radiomic features", Computers in Biology and Medicine, Vol. 174, 2024, p. 108389.
[11] Brown T. A., "Confirmatory factor analysis for applied research", New York, Guilford Publications, 2015.
[12] Mann H. B., Whitney D. R., "On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other", The Annals of Mathematical Statistics, 18, 1, 1947, pp. 50–60.
[13] Wilcoxon F., "Individual Comparisons by Ranking Methods", Biometrics Bulletin, 1, 6, 1945, pp. 80–83.
[14] Kerby D. S., "The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation", Comprehensive Psychology, 3, 2014, pp. 1–9.
[15] Cohen J., Statistical Power Analysis for the Behavioral Sciences, 2nd ed., Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.
[16] JASP Team, "JASP (Version 0.18.3)", Computer software, 2024.