Territorial Fairness in Large-Scale Academic Risk Prediction: Comparing National and State-Level Machine Learning Models in Brazil

Tobias Vieira Francisco[1,2], Abílio Nogueira Barros[2], Felipe Vieira Roque[2], Tiago Paulino[2], Augusto Schmidt[2], Flavia Galvani[3], Rafael Oliveira[2,5], Leonardo Brandão Marques[2], Diego Dermeval[2], Pedro Barreto[4], Anita Gea Martinez Stefani[4], Marisa de Santana da Costa[4], Emanuel Marques Queiroga[2,5], Elthon Oliveira[2], Ig Ibert Bittencourt[2], Cristian Cechinel[2,6]*, and Thales Vieira[2]

[1] Instituto Federal de Educação, Ciência e Tecnologia Sul-Rio-Grandense (IFSul), Brazil
tobiasfrancisco@ifsul.edu.br

[2] Núcleo de Excelência em Tecnologias Sociais (NEES), Universidade Federal de Alagoas (UFAL), Brazil
{abilio.barros, felipe.roque, tiago.paulino, augusto.schmidt, rafael.oliveira, leonardo.marques, diego.matos, elthon.oliveira, ig.ibert, thales.vieira}@nees.ufal.br

[3] Blavatnik School of Government, University of Oxford, United Kingdom
flavia.galvani@bsg.ox.ac.uk

[4] Ministério da Educação (MEC), Brasília, Brasil
{pedrobarreto, anitastefani, marisacosta}@mec.gov.br

[5] Universidade Tecnológica Federal do Paraná (UTFPR), Brazil
emanuelmqueiroga@gmail.com

[6] Centro de Ciências, Tecnologias e Saúde, Universidade Federal de Santa Catarina, Araranguá, Brazil
cristian.cechinel@ufsc.br

Abstract. Early identification of students at academic risk is a central challenge for large and decentralized educational systems. In countries such as Brazil, pronounced regional disparities raise concerns not only about predictive performance, but also about whether machine learning models generalize equitably across territories. This study examines the role of regionalization in academic risk prediction by comparing national and state-level models trained under a unified experimental pipeline using longitudinal administrative data from over six million upper secondary student enrollments across all Brazilian states. Multiple supervised learning algorithms are evaluated, with Random Forest selected for detailed analysis due to its robust overall performance. Territorial fairness is assessed through an operationalization of Equal Opportunity and Equalized Odds based on state-level true and false positive rates. Results show that while national and state-level models achieve similar aggregate performance, substantial disparities persist in Recall across

* Corresponding author

states. State-level models improve local risk detection in a small subset of states, often at the cost of increased false positives. These findings indicate that regional specialization is not uniformly beneficial and should be understood as a context-dependent trade-off between improved local detection and governance complexity. By separating territorial fairness auditing from performance-based model comparison, this study provides an evidence-based framework for reasoning about regionalization in large-scale educational risk prediction.

Keywords: Learning Analytics · Academic Risk Prediction · Algorithmic Fairness · Territorial Equity

1 Introduction

Upper secondary school non-completion, including dropout, grade repetition, and chronic absenteeism, remains a persistent challenge for educational systems worldwide, with long-term social and economic consequences for individuals and societies [20]. In Brazil, this phenomenon is particularly pronounced, reflecting structural inequalities, regional disparities, and institutional constraints that limit the effectiveness of universal educational policies [17]. Prior research has shown that school disengagement is a cumulative process shaped by academic, behavioral, family, and contextual factors that often emerge early in students' educational trajectories [15]. In response, Learning Analytics (LA) and Educational Data Mining (EDM) have gained prominence as approaches for supporting data-driven decision-making in education [18]. By leveraging administrative, behavioral, and academic data, machine learning models have demonstrated strong potential for early identification of students at risk of adverse academic outcomes, including dropout and failure [11]. Large-scale studies in national contexts such as Denmark, Uruguay, and Brazil have shown that such models can be trained using routinely collected data, achieving meaningful predictive performance even under severe class imbalance [16].

Despite these advances, much of the existing evidence is derived from institution- or district-level analyses that implicitly assume relatively homogeneous educational contexts [6]. In contrast, national education systems—particularly in large and socioeconomically diverse countries—exhibit substantial territorial heterogeneity in school infrastructure, student profiles, and policy implementation [8]. In Brazil, historical, economic, and administrative differences across states are reflected in both data distributions and model behavior, raising concerns that a single nationwide model may obscure local patterns and yield uneven performance across regions [17]. The use of machine learning in education also raises critical questions related to fairness, equity, and responsible data use [13]. Recent work has shown that models trained on aggregated data can inadvertently reinforce existing inequalities when disparities in predictive errors across groups or contexts are not explicitly examined [3]. These concerns are especially salient in public education systems, where predictive models increasingly inform policy decisions and the allocation of limited resources [17].

In Brazil, large-scale educational data infrastructures such as the Sistema Gestão Presente (SGP) [2] have enabled the integration of administrative and behavioral data at a national scale. This infrastructure plays a central role in supporting public policies such as the *Pé-de-Meia* program, a conditional cash transfer initiative aimed at reducing dropout, which relies on these data to monitor student participation and enforce eligibility criteria. Beyond policy enforcement, the availability of such integrated datasets creates opportunities for the development of predictive models capable of identifying students at risk of adverse academic outcomes at early stages. This, in turn, supports more targeted and data-driven intervention strategies, aligning learning analytics approaches with large-scale policy implementation in highly heterogeneous educational systems.

In this paper, we examine the generalization behavior of machine learning models for academic risk prediction in a territorially heterogeneous educational system. We adopt a two-scale modeling strategy, training a national-level model using data from all Brazilian states and state-level models using the same pipeline restricted to individual states. Using a fairness-oriented diagnostic perspective, we operationalize Equal Opportunity and Equalized Odds through state-wise True Positive and False Positive Rates to assess whether the national model generalizes uniformly across regions. This framework supports a structured analysis of territorial disparities and informs a performance–governance trade-off analysis, identifying when state-level specialization may offer meaningful advantages beyond aggregate predictive performance. The remainder of this paper is structured as follows. Section 2 reviews related work on predictive modeling in education. Section 3 describes the experimental design and evaluation framework. Section 4 presents and discusses the empirical findings, and Section 5 concludes the paper with final remarks and directions for future work.

2 Related Work

The use of Learning Analytics (LA) and Educational Data Mining (EDM) to predict academic risk and school dropout has grown substantially over the past decade. Early contributions established the role of analytics in supporting evidence-based and learning-centered decision-making in education [7], while subsequent surveys documented the widespread adoption of machine learning techniques for predicting academic performance, failure, and dropout across secondary and higher education settings [1].

Several large-scale studies have demonstrated the feasibility of applying predictive models at national or multi-district levels using administrative educational data. In Uruguay, Queiroga et al. [16] reported a nationwide learning analytics initiative in K–12 education, employing Random Forest models to predict grade repetition and dropout using data from more than 258,000 students. Similarly, Sara et al. [19] conducted a national study in Denmark using registry data from over 36,000 upper secondary students, achieving high predictive performance with Random Forest models. In the United States, Christie et al. [5] described a machine learning–based early warning system deployed across 32 states, covering

millions of educational records and integrating XGBoost models directly into school management platforms. In South Korea, Lee and Chung [12] analyzed national data from more than 165,000 students, showing that tree-based and boosting methods outperformed alternative approaches for dropout prediction, particularly under severe class imbalance.

Beyond technical feasibility, prior work has emphasized the importance of aligning predictive modeling with educational decision-making processes. Lakkaraju et al. [11], in collaboration with large U.S. school districts, highlighted that commonly reported performance metrics such as AUC and accuracy may not fully capture the operational needs of educators, underscoring the relevance of selecting evaluation criteria that reflect intervention priorities.

Despite these advances, important gaps remain in the literature. In particular, systematic analyses comparing centralized (e.g., national-level) and decentralized (e.g., state- or district-level) modeling strategies are still limited, especially in contexts characterized by strong territorial heterogeneity [6]. Moreover, while large-scale studies often report aggregate performance, they rarely examine how predictive errors are distributed across regions or consider fairness-related implications of model generalization. Addressing these gaps is essential for informing governance-aware modeling strategies in decentralized educational systems.

3 Methodology

This section describes the materials and methods employed to analyze the generalization behavior of machine learning models for academic risk prediction in a territorially heterogeneous context. The methodology is organized into three main components: (A) data collection, (B) preprocessing and modeling, and (C) evaluation and analysis. Figure 1 summarizes the overall experimental framework.

3.1 Data Collection, Feature Selection, and Preparation

The construction of the dataset (see Table 1 for variable names and descriptions) was based on the consolidation of two main sources of information: the Sistema de Gestão Presente and the 2024 Basic Education Census. The SGP operates as the Basic Education Data Platform, integrating administrative and behavioral data from multiple education networks to provide a unified view of students' educational trajectories. As a centralized infrastructure, it supports large-scale monitoring initiatives in Brazilian public education. In this study, it serves as the primary source for constructing the analytical dataset used in model development.

The dataset comprises attendance records associated with high school enrollments, covering the first, second, and third grades, reported monthly throughout the 2024 academic year, with the exception of January. The data were extracted directly from the SGP, the system responsible for collecting and consolidating information on high school students across the national territory [2]. Records are organized at the enrollment level and include students regularly enrolled in the Brazilian state school network in 2024, totaling 6,170,183 enrollment records.

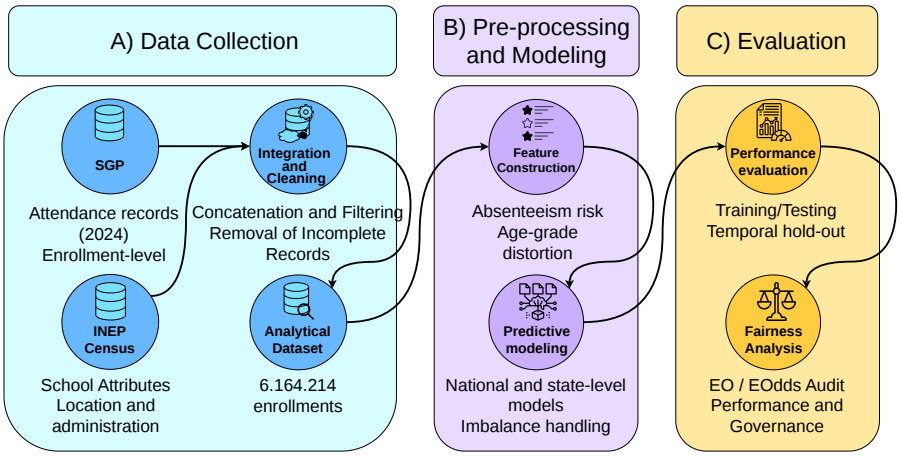

Fig. 1. Overview of the proposed approach, comprising data collection, preprocessing and modeling, and evaluation stages.

We aggregate enrollment attendance data at the school level to compute absenteeism metrics, defined as sustained student absence over the academic year. The proportions of students in each absenteeism level follow the framework proposed by [10]: Low ($> 95\%$), indicating high attendance; Moderate (91–95%), indicating few absences; Significant (81–90%), reflecting frequent absence; High (71–80%), indicating a concerning pattern; and Severe ($\leq 70\%$), representing a high risk of school dropout. Based on these proportions, variables were created to describe each school's student composition according to the adopted absenteeism criteria. Using these measures, a derived feature was created (`entity_risk_abs`) to indicate overall school-level absenteeism risk. Schools with more than 50% of enrollments classified as Low or Moderate were labeled as class 1, whereas schools with more than 50% of enrollments classified as Significant, High, or Severe were labeled as class 0, characterizing high absenteeism levels. A set of features incorporated into the dataset refers to the Age–Grade Distortion Rate (TDI). A student is considered to be in an age–grade distortion condition when they are two years or more above the expected age for the grade attended [9]. Based on this criterion, three features were created: `in_tdi`, `pp_tdi_serie`, and `pp_mat_tdi`. Information related to school location, including region, federative unit, and location type (urban or rural), was obtained from the 2024 Basic Education Census published by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). After integrating all data sources through dataset concatenation, 5,969 records were removed due to incompleteness, corresponding to approximately 0.10% of the original dataset.

The final dataset comprises a total of 6,164,214 enrollments. Of these, 623,901 enrollments (10.12%) correspond to students associated with multiple enrollments (`multiple_enrollments`) during the academic year. Among them, 623,450 enrollments (10.11%) reported at least one occurrence in which enrollments as-

Table 1. Description of input features used in the model

Feature Name	Description
attendance_reported_{month}	Indicates whether attendance was reported in the given month.
hours_offered_{month}_2024	Total instructional hours offered in the given month.
hours_attended_{month}_2024	Total instructional hours attended in the given month.
attendance_rate_{month}_2024	Monthly attendance rate.
multi_enroll	Indicates student multiple active enrollments.
multi_enroll_with_attendance	Indicates whether another enrollment reported attendance.
enroll_termination_reason	Reason for enrollment termination.
entity_{school_location}	School location (urban or rural).
student_in_{year}_year	Grade level in upper secondary education.
entity_{edu_gov_level}	School adm. level (municipal/state/federal).
entity_risk_abs	Indicates high absenteeism at the school level.
in_tdi	Indicates age–grade distortion.
pp_tdi_serie	Percentage of students with age–grade distortion.
pp_mat_tdi	Percentage of enrollments with age-grade distortion in the school.
no_region_entity_{region}	School macro-region indicators.
no_uf_entity_{uf}	State (federative unit) indicators.

sociated with the same student recorded attendance within the same month (`multiple_enrollments_with_attendance`). Regarding administrative dependency, the dataset is predominantly composed of enrollments from state-administered schools (`entity_state`), totaling 6,129,921 enrollments (99.44%). Municipal schools (`entity_municipal`) account for 25,101 enrollments (0.41%), while federal schools (`entity_federal`) represent 9,192 enrollments (0.15%). Concerning school location, 5,814,378 enrollments (94.33%) are associated with urban schools (`entity_loc_urbana`), whereas 349,836 enrollments (5.67%) correspond to rural schools (`entity_loc_rural`).

The distribution of enrollments across upper secondary grades indicates that 2,391,491 enrollments (38.78%) correspond to students enrolled in the first year (`student_in_1_year`), 2,017,858 enrollments (32.74%) in the second year (`student_in_2_year`), and 1,754,865 enrollments (28.48%) in the third year (`student_in_3_year`).

The target variable was defined at the enrollment level using administrative information that captures the student's enrollment status at the end of the academic year, as recorded in the SGP. This variable reflects the officially reported reason for enrollment termination and serves as a proxy for academic risk within the public education system. Prior to label construction, records associated with administrative inconsistencies and cases in which the final status indicated death were excluded, as these situations are not informative for modeling educational risk. The remaining enrollment outcomes were then grouped according to whether they reflected successful academic progression or trajectories associated with potential vulnerability. Outcomes corresponding to approval or confirmed completion, either through regular progression or alternative certification pathways, were labeled as not at risk (class 0). All other outcomes, including dropout, abandonment, failure, transfers, suspensions, and reclassification events, were labeled as at risk (class 1), as they represent conditions in which continued educational engagement is disrupted or uncertain.

3.2 Modeling and Evaluation

Five algorithms were selected: Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), AdaBoost (AB), and Multilayer Perceptron (MLP). Logistic Regression was adopted as the baseline model, as it represents a low-complexity linear approach widely recognized for its simplicity and interpretability [14]. In contrast, the MLP was included as an approximate upper bound of performance due to its higher expressiveness and ability to model complex nonlinear relationships, which are often associated with predictive gains in large-scale educational settings [11]. The dataset was partitioned into training and testing subsets using a stratified split with a 90%/10% ratio. Within the training set, five-fold cross-validation was applied using the `cross_validate` function, evaluating the metrics Accuracy, Precision, Recall, F1-score, and AUC-ROC.

As a methodological choice, no systematic hyperparameter optimization techniques were employed. This decision reflects the comparative nature of the study, whose primary objective is not to maximize the absolute performance of a specific model, but to isolate the effect of territorial segmentation by ensuring that both nationwide and state-level models are evaluated under identical configurations. Using the same pipeline and parameters across all scenarios enables a controlled comparison, in which observed differences in performance and fairness can be attributed to the modeling strategy rather than to differences in optimization or calibration.

Random Under Sampling (RUS) was applied exclusively to the training data. This approach is expected to introduce less distortion to the data characteristics. RUS was performed independently for each state-level subset in order to preserve local class balance and reflect regional data distributions. However, in states with smaller student populations, state-level RUS may attenuate rare-pattern signals associated with high-risk cases, potentially affecting local prevalence estimation and calibration, with possible implications for fairness-related outcomes. This trade-off represents a known threat to validity and is partially mitigated by the consistent application of the same resampling strategy across all models and by the comparative focus of the analysis.

Model training and evaluation were conducted longitudinally through eleven independent training cycles, in which all algorithms were trained on cumulative datasets defined by monthly cutoffs of school attendance records (from February to December). This design simulates the progressive accumulation of information over the academic year, allowing analysis of how predictive performance evolves as additional behavioral evidence becomes available [16]. After identifying the algorithm with the highest average performance at the national level, a second experimental phase applied the same classifier to train state-specific models using datasets derived from the national training and test partitions, ensuring methodological consistency and preventing data leakage. The performance of these state-specific models was then compared against that of the national model when evaluated on the corresponding state-level subsets, enabling a controlled assessment of the effects of regional specialization in a territorially and socioeconomically heterogeneous educational system [5].

4 Results and Discussion

This section presents and discusses the results of the proposed predictive models for identifying students at risk. We first report the performance of the national-level model trained using data from all Brazilian federative units. Next, we analyze the results of state-level models, highlighting differences in predictive performance across regions. Finally, we evaluate the generalization ability of the national and state models by examining their cross-level applicability and robustness. The findings are discussed regarding the implications for large-scale educational monitoring and early identification of at-risk students.

4.1 Overall Predictive Performance

First, we present a comparative analysis of five supervised classification algorithms, trained on the national-level dataset to predict students at academic risk. Random Forest achieved the most robust and balanced performance across all evaluated metrics, including Accuracy, F1-score, and AUC-ROC, while maintaining lower variance across cross-validation folds. As illustrated in Figure 2, RF showed consistently strong performance across the academic year, with improvements becoming increasingly evident as more monthly data accumulated.

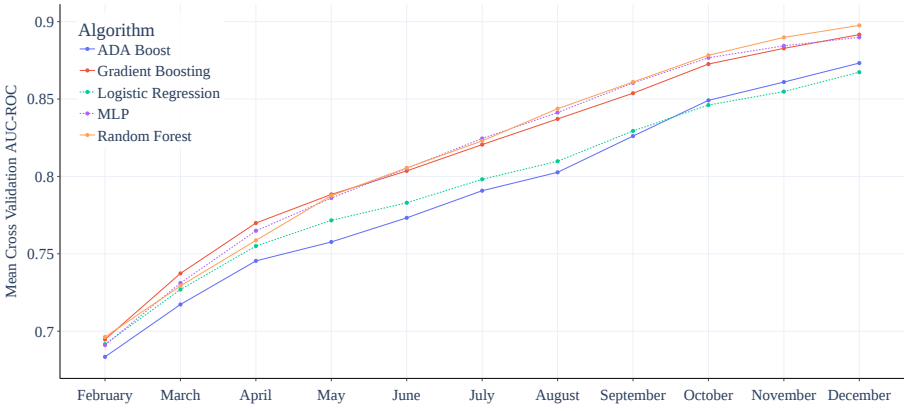

Fig. 2. Longitudinal Mean AUC-ROC performance by algorithm at the national level

After selecting RF as the best-performing algorithm at the national level, a second experimental stage evaluated its behavior under regional specialization. In this phase, the national RF model was applied to state-level test datasets and compared against models trained exclusively with data from each respective federative unit. Across all months and experimental folds, the national and state-level models exhibited very similar overall performance. Aggregate metrics such

as AUC, F1-score, and Recall showed no consistent dominance of one modeling strategy over the other when averaged across states. This proximity in global performance indicates that, from a purely predictive standpoint, both approaches are viable and that performance differences are subtle.

Nonetheless, a closer inspection reveals variation across states and metrics. While AUC and F1-score remain relatively stable, Recall presents larger dispersion between national and state-specific implementations (as can be seen in Fig. 3). Given that Recall directly reflects the model's ability to correctly identify students at risk, these variations warrant deeper investigation beyond aggregate performance summaries, especially in early warning system (EWS) contexts [12].

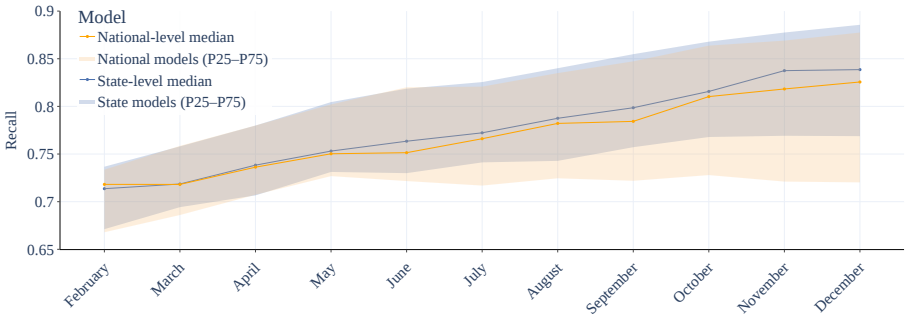

Fig. 3. Median Recall over time for the national and state-level models. The shaded region indicates the range of Recall values across states.

4.2 Territorial Fairness: Operationalizing Equal Opportunity

We evaluate the generalization behavior of national and state-level predictive models through a fairness-oriented analysis of EO across Brazilian states. Rather than adopting EO and Equalized Odds (EOdds) in their classical group-based fairness formulations, this study operationalizes a notion of *territorial fairness*. In this context, states are treated as comparison groups, and fairness is assessed through disparities in error rates across territories.

EO is operationalized through the True Positive Rate (TPR) computed independently for each state. This choice aligns with the study's prioritization of Recall and reflects the normative importance of ensuring that students with similar risk profiles have comparable probabilities of being correctly identified, regardless of their state of residence [17]. Disparity in EO is quantified as the difference between the maximum and minimum state-level true positive rates, as defined in $\Delta_{EO} = \max_s (TPR_s) - \min_s (TPR_s)$, where TPR_s denotes the true positive rate observed in state s. The results indicate a substantial territorial gap. The observed ΔEO reaches approximately 0.26, meaning that the probability

of correctly identifying at-risk students differs by up to 26 percentage points between states. In practical terms, disparities of this magnitude indicate that students in some states are substantially less likely to be correctly identified as at risk compared to others.

4.3 Equalized Odds and False Positive Trade-offs

To complement the EO analysis, EOdds is examined by jointly considering TPR and False Positive Rate (FPR). Consistent with the extended evaluation pipeline, EOdds is summarized using the average odds disparity, defined as $\Delta EOdds = \frac{\Delta TPR + \Delta FPR}{2}$. Post-processing results show a ΔFPR of approximately 0.20 and an average $\Delta EOdds$ of roughly 0.23. These values indicate that disparities are not confined to missed detections but also involve uneven exposure to false alarms across states. Importantly, this reinforces that fairness considerations cannot be reduced to Recall alone, even when Recall is the prioritized metric.

4.4 Investigating the Sources of Territorial Disparity

To assess whether observed EO disparities could be explained by structural characteristics of the data, statistical analyses were conducted examining the relationship between ΔEO and two factors: (i) state-level risk prevalence and (ii) sample size. To test (i), we compute the Spearman rank correlation between state-level TPR and the proportion of students in each state who are truly at academic risk. The association between EO disparity and the prevalence of risk was weak and did not reach statistical significance (Spearman's $\rho = 0.229$, $p = 0.25$). This suggests that differences in underlying risk distribution alone do not account for the observed territorial gaps in TPR.

In a similar way, the association between EO disparity and sample size (ii) was tested by applying the Spearman rank correlation between state-level TPR and the number of at-risk students in states. In contrast to the previous finding, the presented correlation was moderate and statistically significant (Spearman's $\rho = 0.434$, $p = 0.023$). While this result indicates that data volume may influence fairness outcomes, it does not fully explain the magnitude or direction of disparities. States with larger samples do not systematically achieve either higher or lower fairness gaps, suggesting that additional contextual or structural factors are at play.

4.5 Analyzing Performance-Governance Trade-offs Across Models

To further investigate whether territorial disparities identified in the fairness audit could be mitigated through regional specialization, we performed a paired comparison between national and state-level models at the state level. This analysis focuses on differences in True Positive Rate (TPR, corresponding to Recall for the positive class) and False Positive Rate (FPR), computed as state–national differences and evaluated using the Wilcoxon signed-rank test applied to monthly

hold-out Recall values. While this step does not constitute a direct fairness assessment, it examines whether state-level models yield statistically robust gains in risk detection relative to the national model, given the additional analytical and governance complexity associated with specialization.

Table 2. Operational model selection under recall–false-alarm trade-offs

State	National TPR	State-level TPR	ΔTPR	ΔFPR	Recall p-value	Decision
Acre	0.7011	0.7193	0.0182	0.0125	0.0049	No diff
Alagoas	0.7661	0.7713	0.0052	0.0108	0.0312	No diff
Amapá	0.7152	0.7168	0.0016	-0.0150	0.2754	No diff
Amazonas	0.8578	0.8484	-0.0094	-0.0118	0.4785	No diff
Bahia	0.7543	0.7607	0.0065	-0.0011	0.0010	No diff
Ceará	0.8555	0.8389	-0.0165	-0.0235	0.0010	No diff
Espírito Santo	0.8665	0.8555	-0.0110	-0.0148	0.0420	No diff
Federal District	0.8266	0.8292	0.0025	-0.0002	0.6152	No diff
Goiás	0.7876	0.7853	-0.0022	-0.0005	0.8262	No diff
Maranhão	0.7536	0.7631	0.0095	0.0069	0.0010	No diff
Mato Grosso	0.8488	0.8397	-0.0091	-0.0095	0.0049	No diff
Mato Grosso do Sul	0.8271	0.8322	0.0051	0.0027	0.0010	No diff
Minas Gerais	0.8174	0.8212	0.0038	0.0019	0.0098	No diff
Pará	0.6965	0.7188	0.0222	-0.0104	0.0029	State-level model
Paraíba	0.7630	0.7626	-0.0003	-0.0049	0.0137	No diff
Paraná	0.8215	0.8219	0.0004	0.0015	0.0059	No diff
Pernambuco	0.7782	0.7813	0.0031	-0.0028	0.0400	No diff
Piauí	0.7378	0.7516	0.0138	0.0012	0.0049	No diff
Rio de Janeiro	0.7305	0.7309	0.0004	0.0030	0.1416	No diff
Rio Grande do Norte	0.7970	0.7917	-0.0053	-0.0005	0.0098	No diff
Rio Grande do Sul	0.7799	0.7796	-0.0004	0.0001	0.7480	No diff
Rondônia	0.7260	0.7706	0.0445	0.0221	0.0049	State-level model
Roraima	0.6228	0.6267	0.0039	0.0139	0.0312	No diff
Santa Catarina	0.6097	0.6325	0.0228	0.0226	0.0029	State-level model
São Paulo	0.8475	0.8465	-0.0010	-0.0020	0.0049	No diff
Sergipe	0.7040	0.7634	0.0594	0.0435	0.0010	State-level model
Tocantins	0.6817	0.6877	0.0061	0.0052	0.0020	No diff

As shown in Table 2, only a small subset of states exhibits statistically significant improvements in Recall when using state-level models, with gains ranging from approximately 2 to 6 percentage points. However, these improvements are frequently accompanied by increases in False Positive Rate (FPR), making explicit the trade-off between enhanced risk detection and higher false-alarm rates. To complement the tabular comparison, Figure 4 visualizes these differences by positioning states according to ΔTPR and ΔFPR (state minus national). Shaded regions, defined empirically using the median absolute deviation, delineate zones of meaningful advantage and trade-off, providing an operational perspective on when regional specialization yields balanced gains and when it imposes additional governance costs. For most states, differences between national and state-level models are not statistically significant, and in several cases the national model slightly outperforms its specialized counterpart. These results indicate that regional specialization should not be interpreted as a uniformly superior modeling strategy. Instead, its advantages are context-dependent and must be evaluated jointly under EO and EOdds criteria, consistent with recent discussions on equitable ML deployment in education [13].

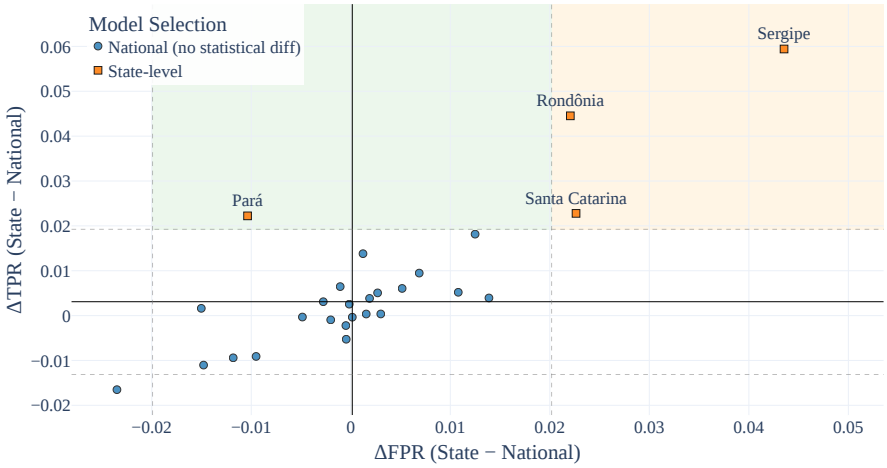

Fig. 4. Performance–governance trade-offs between national and state-level models.

4.6 Implications for Model Selection and Governance

The results suggest that there is no single modeling approach that can be considered best across the majority of states. Although national and state-level models achieve comparable overall performance, substantial territorial disparities persist, particularly when fairness is evaluated through Recall-based EO.

These disparities are not fully explained by sample size or risk prevalence, underscoring that fairness-aware governance requires analyses that go beyond conventional performance summaries.In this context, EO and EOdds serve as **diagnostic tools** to audit territorial generalization, while the national-versus-state comparison supports an **operational decision framework**. Rather than selecting models solely based on average performance, the proposed approach prioritizes Recall—reflecting the educational imperative to minimize missed detections of students at risk—while explicitly accounting for the operational costs associated with increased false positives [4,16]. This distinction clarifies that fairness considerations guide *where to look*, whereas deployment decisions determine *when to intervene*.

From a policy and implementation perspective, these findings suggest that centralized models can serve as a stable default in large-scale educational systems, but should be complemented by targeted regional specialization when fairness diagnostics and statistical evidence jointly indicate clear benefits. By separating territorial fairness auditing from performance-driven model selection, this study reframes model deployment as a governance problem that balances equity, effectiveness, and feasibility in highly heterogeneous educational contexts.

When contrasted with prior work on large-scale dropout prediction and early warning systems, our results reinforce a recurring observation in the literature:

predictive performance alone is often insufficient to support deployment decisions in heterogeneous educational systems. Large-scale studies at national or multi-district levels have demonstrated the effectiveness of centralized predictive models, typically evaluated through aggregate metrics such as AUC and F1-score [11,12,5]. However, despite acknowledging contextual and institutional heterogeneity, these studies primarily report average performance and provide limited analysis of how predictive errors are distributed across territorial units or regions [19]. This limitation becomes especially salient in countries with strong regional diversity and large geographic extent, such as Brazil, where socioeconomic conditions, school infrastructures, and student trajectories vary substantially across states [17]. Such structural differences are reflected in the variability and specificity of the data itself, directly impacting model behavior across regions. While regional analyses have been explored in other national contexts [8], these studies primarily focus on uncovering local patterns or improving predictive accuracy, without explicitly framing regional disparities as a fairness problem.

In contrast, this study advances the state of the art by explicitly integrating territorial fairness into the model selection process at an unprecedented national scale. By operationalizing EO and EOdds through state-level TPR and FPR, we assess whether a national model generalizes equitably across regions, rather than merely whether it performs well on average. This approach moves beyond the binary question of whether state-level models outperform a national one and instead provides evidence-based criteria to determine *when* specialization may be analytically justified. In doing so, model selection is reframed as a fairness-sensitive assessment that accounts for territorial heterogeneity, rather than a purely predictive optimization task.

5 Final remarks

This study examined the behavior of large-scale machine learning models for academic risk prediction in a territorially heterogeneous educational system. By comparing national and state-level modeling strategies under a unified experimental pipeline, we showed that similar aggregate performance can conceal substantial regional variability in model behavior. The adoption of a territorial fairness perspective, operationalized through Equal Opportunity and Equalized Odds based on Recall, enabled a systematic assessment of how well a national model generalizes across states, reframing fairness as a governance-relevant property, rather than as a classical protected-attribute constraint.

Despite practical and methodological challenges inherent to this type of analysis, such as the country's continental scale, pronounced regional diversity, highly variable data distributions, and the computational constraints of large-scale experimentation; the proposed framework demonstrates that fairness-oriented diagnostics can be effectively integrated into model comparison. Even in the absence of fine-grained contextual variables and under limited validation flexibility, the combination of territorial fairness auditing and governance-aware performance analysis provides a principled way to reason about model selection in decentralized

educational systems, moving beyond aggregate accuracy toward more context-sensitive and equitable analytical decisions.

Future work should explore fairness-aware modeling strategies based on hybrid and cascading architectures that combine national models with progressively localized components, potentially extending specialization beyond the state level to intra-state regions. Additional efforts should assess the role of class balancing techniques in mitigating territorial disparities. Finally, longitudinal analyses of fairness indicators remain necessary to evaluate the stability of territorial gaps over time and their responsiveness to policy or behavioral changes.

Acknowledgments

This work was supported by the Brazilian Ministry of Education (MEC) under grants TED13914 and TED11476; by the Brazilian National Council for Scientific and Technological Development (CNPq) under grant No. 445016/2024-8; by the project Ia.Edu – National Institute of Artificial Intelligence in Unplugged Education (grant No. 4084883/2024-5); and by the Research Support Foundation of the State of Alagoas (FAPEAL) under grant No. FAPEAL60030.0000001481/2025.

Disclosure of Interests

The authors have no competing interests to declare.

References

1. Balaji, P., Alelyani, S., Qahmash, A., Mohana, M.: Contributions of machine learning models towards student academic performance prediction: a systematic review. Applied Sciences **11**(21), 10007 (2021)
2. Barros, A.N., Queiroga, E.M., Marcolino, M.R., Silva, D.B.L., Dermeval, D., Lima, A., Marques, L.B., Cechinel, C., Vieira, T.: Ensuring data quality in national educational databases: Insights from brazil's centralized database of high school students' data. In: Conference on Digital Government Research. vol. 1 (2025)
3. Bulut, O., Wongvorachan, T., He, S., Lee, S.: Enhancing high-school dropout identification: A collaborative approach integrating human and machine insights. Discover Education **3**(1), 109 (2024)
4. Christenson, S.L., Thurlow, M.L.: School dropouts: Prevention considerations, interventions, and challenges. Current Directions in Psychological Science **13**(1), 36–39 (2004)
5. Christie, S.T., Shi, A., Gobert, L., Rosen, C.: Machine-learned school dropout early warning at scale. In: Proceedings of the 12th International Conference on Educational Data Mining (2019)
6. Coleman, C., Baker, T., Stephenson, E.: A better cold-start for early prediction of student at-risk status in new school districts. In: Proceedings of the 12th International Conference on Educational Data Mining (2019)
7. Ferguson, R.: Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning **4**(5/6), 304–317 (2012)

8. Hernández-Leal, E., Duque-Méndez, N.D., Cechinel, C.: Unveiling educational patterns at a regional level in colombia: data from elementary and public high school institutions. Heliyon **7**(9) (2021)
9. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira: Dicionário de indicadores educacionais: fórmulas de cálculo. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília (2004), https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/dicionario_de_indicadores_educacionais_formulas_de_calculo.pdf
10. Kearney, C.A.: Integrating systemic and analytic approaches to school attendance problems: Synergistic frameworks for research and policy directions. In: Child & Youth Care Forum. vol. 50, pp. 701–742. Springer (2021)
11. Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., Addison, K.L.: A machine learning framework to identify students at risk of adverse academic outcomes. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1909–1918 (2015). https://doi.org/10.1145/2783258.2788620
12. Lee, S., Chung, J.Y.: The machine learning-based dropout early warning system for improving the performance of dropout prediction. Applied Sciences **9**(15), 3093 (2019). https://doi.org/10.3390/APP9153093
13. Mangal, M., Pardos, Z.A.: Implementing equitable and intersectionality-aware ml in education: A practical guide. British Journal of Educational Technology **55**(5), 2003–2038 (2024)
14. Márquez-Vera, C., Cano, A., Romero, C., Noaman, A.Y.M., Fardoun, H.M., Ventura, S.: Early dropout prediction using data mining: a case study with high school students. Expert Systems **33**(1), 107–124 (2016)
15. Parr, A.K., Bonitz, V.S.: Role of family background, student behaviors, and school-related beliefs in predicting high school dropout. The Journal of Educational Research **108**(6), 504–514 (2015)
16. Queiroga, E.M., Paragarino, V.R., Casas, A.P., Primo, T.T., Munoz, R., Ramos, V.C., Cechinel, C.: Early prediction of at-risk students in secondary education: A countrywide k-12 learning analytics initiative in uruguay. Information **13**(9), 401 (2022). https://doi.org/10.3390/info13090401
17. Queiroga, E.M., Siqueira, E.S., Portela, C.D.S., Cordeiro, T.D., Bittencourt, I.I., Isotani, S., Mello, R.F., Muñoz, R., Cechinel, C.: Data-driven strategies for achieving school equity: Insights from brazil and policy recommendations. IEEE Access **12**, 101646–101659 (2024)
18. Romero, C., Ventura, S.: Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery **10**(3), e1355 (2020)
19. Sara, N.B., Halland, R., Igel, I.G., Alstrup, S.: High-school dropout prediction using machine learning: A danish large-scale study. In: Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (2015)
20. UIS — UNESCO Institute for Statistics: Global education digest 2012—opportunities lost: The impact of grade repetition and early school leaving (2012)