From Importance to Action: Evaluating Counterfactual Explanations under Constraints in Brazilian Public Schools

Ana Beatriz P. de A. Rebouças [1], **César L. C. Mattos**[2]

[1]Department of Statistics and Applied Mathematics

[2]Department of Computing
Federal University of Ceará (UFC), Fortaleza – CE

anareboucas@alu.ufc.br, cesarlincoln@dc.ufc.br

Abstract. *Machine learning has been increasingly used to predict educational outcomes. However, prediction alone is insufficient for meaningful intervention, where the main goal is not to label students or schools, but to identify actionable paths toward improvement. Counterfactual Explanations (CFEs) address this by identifying minimal feature changes associated with better outcomes. In large-scale exams such as Enem, the student cohort changes each year, limiting individual-level analysis, while municipality-level aggregation introduces substantial heterogeneity. Hence, we analyze CFEs at school level, combining Enem 2024 results with the 2024 Brazilian School Census. Focusing on a realistic policy setting, we restrict the analysis to public state schools and incorporate intervention constraints through different feature mutability levels. First, we train five classifiers to predict school performance. Then, we generate and evaluate CFEs for the best-performing model (XGBoost). We compare results across mutability constraint levels, generation methods (DiCE, Growing Spheres, and FACE), and school subgroups. Our results reveal actionability discrepancies across subgroups under operational constraints, with implications for equitable interventions. We also contrast SHAP feature importance with CFEs, revealing a gap between predictive importance and actionable change, which is critical for distinguishing explanatory relevance from intervention-oriented insights.*

1. Introduction

Advances in data availability, computational power, and model flexibility have enabled the widespread adoption of machine learning (ML) to support decisions in diverse and high-stakes domains. However, it has also raised concerns related to fairness, accountability, transparency, and ethics (FATE) [Verma et al. 2024]. In this context, eXplainable Artificial Intelligence (XAI) emerged as a key requirement, with the aim of making model behavior more understandable and trustworthy [Khosravi et al. 2022].

For education applications, where dropout prediction and performance forecasting are common learning tasks [Buñay-Guisñan et al. 2026], works applying XAI usually focus on identifying the most relevant features influencing model decisions [Silva Filho et al. 2023, Cavus and Kuzilek 2024]. In this domain, the model prediction is often insufficient, since, by itself, it cannot be seen as deterministic, but rather as an indication of what could happen if nothing changes. As stated by [Ustun et al. 2019], "individuals should have the ability to change the decision of the model". Otherwise, they "will lack agency over a decision that affects their livelihood".

Although XAI improves transparency and model interpretation, most techniques, such as SHAP (SHapley Additive exPlanations) [Lundberg and Lee 2017] and LIME (Local Interpretable Model-agnostic Explanations) [Ribeiro et al. 2016], remain descriptive, focusing on why a prediction was made rather than how to act upon it. On the other hand, Counterfactual Explanations (CFEs) shift the focus from explanation to action, indicating the minimal changes required in the input to achieve a desired outcome.

Despite their potential, applying CFEs in education is challenging, since outcomes are strongly influenced by socioeconomic factors, often identified as highly relevant by predictive models [Dutra et al. 2023, Suaza-Medina et al. 2024]. Improving such conditions is a valid long-term policy goal, but many of these variables are not directly actionable in the short term [Guidotti 2024, Buñay-Guisñan et al. 2026]. This creates a gap between what models identify as important and what can realistically be changed.

Moreover, most existing works on explainability in education aim at the individual student level. However, intervention can also occur at school or municipality levels. In the case of large-scale national exams, such as Enem (Exame Nacional do Ensino Médio, Brazil's national high school exam used for higher education admission), the student population changes each year, limiting individual longitudinal analysis. Conversely, municipality-level aggregation introduces high heterogeneity due to differences among schools. Nevertheless, schools themselves represent a promising unit of analysis, as they have their own administrative and pedagogical structures that can be targeted by interventions. Providing actionable recommendations remains challenging, since schools operate under different constraints, resources, infrastructures, and organizational capacities. Thus, average-based recommendations may not be feasible. Conversely, CFEs offer instance-level guidance [Bayrak and Bach 2024] within each school's specific context.

Generating realistic counterfactuals requires constraints that reflect real-world feasibility. Not all features can be changed, and even mutable features may differ in terms of cost, effort, or time required for intervention [Karimi et al. 2021]. Since state-of-the-art methods cannot guarantee all counterfactual desiderata simultaneously [Guidotti 2024], this work aims to highlight the consequences of modeling choices and shed light on their inherent trade-offs. To bridge the gap from prediction to prescription, we leverage a large-scale dataset, comprised of assessment data and School Census information.

To obtain feasible intervention paths personalized to each school, we propose a framework based on different levels of feature mutability. Quantitative evaluation of CFEs faces significant challenges, including inconsistent terminology and metrics [Bayrak and Bach 2024] and a lack of standardized protocols [Moreira et al. 2025]. Thus, given the scarcity of benchmarks that assess the quality of different CFE methods in the educational domain [Cavus and Kuzilek 2024], we make the following contributions:

- A domain-driven feature mutability framework that establishes realistic intervention spaces for CFE generation;
- A comparative benchmarking of CFE algorithms under different restriction levels, exposing the algorithmic trade-offs between actionability, proximity, and sparsity;
- An empirical demonstration of the divergence between feature attribution (SHAP) and actionable changes (CFEs) under mutability constraints;
- An investigation of CFE actionability across school subgroups, revealing how structural inequalities limit the feasibility of these interventions in practice.

2. Background

CFEs aim to bridge the gap between descriptive diagnosis and actionable recommendation. According to [Guidotti 2024]: "factual explanations aid logical reasoning, counterfactuals add new information to what is known about the facts". It is inherently associated with how humans understand the world and is a concept well established in the social sciences. However, translating it into machine explanations was not immediate. Attempts to alter inputs to change outcomes already appeared through inverse classification and adversarial examples, but the formal introduction of counterfactuals as an explainability tool for ML models only occurred with [Wachter et al. 2017], who framed existing optimization techniques within a human-centered explanatory framework.

Since computing real-world causal rules is notoriously difficult, [Wachter et al. 2017] simply altered inputs to change outputs, ignoring causal relationships between variables. This can lead to suggestions that are meaningless or impossible to execute in practice. To address this limitation, [Karimi et al. 2021] argue that true recourse must respect causality, since changing one feature often affects others. [Ustun et al. 2019] focuses instead on actionability constraints, asking which features can realistically be changed at all, distinguishing immutable and actionable features.

2.1. Generation of counterfactual explanations

CFEs generation can be formulated as an optimization problem [Guidotti 2024]. Given a classifier f and an input instance $\mathbf{x}$ with prediction $y = f(\mathbf{x})$, the goal is to find a counterfactual instance $\mathbf{x}'$ that achieves a desired outcome $y' \neq y$ while remaining as close as possible to the original instance [Verma et al. 2024]:

$$\arg\min_{\mathbf{x}'} \; d(\mathbf{x}, \mathbf{x}') \quad \text{s.t.} \quad f(\mathbf{x}') = y', \tag{1}$$

where $d(\mathbf{x}, \mathbf{x}')$ measures the distance between the original instance and its counterfactual. Since directly enforcing $f(\mathbf{x}') = y'$ is difficult in practice, [Wachter et al. 2017] converted it into an unconstrained loss function via Lagrangian relaxation: $\arg\min_{\mathbf{x}'} \max_{\lambda} \; \lambda\big(f(\mathbf{x}') - y'\big)^2 + d(\mathbf{x}, \mathbf{x}')$, where λ balances validity (reaching the desired class) against proximity (staying close to the original). Some strategies treat λ as a tunable (or fixed) hyperparameter; more generally, one can extend this formulation by introducing multiple weighted terms to penalize different desiderata, such as proximity, diversity, or robustness. The result is a family of loss functions: $\mathcal{L} = \lambda_1 \cdot \text{Validity} + \text{Proximity} + \sum_{i=2}^{n} \lambda_i \cdot \text{Desideratum}_i$, where λ_i controls the weight of the corresponding term. However, this leads to complex optimization trade-offs, since satisfying one property often compromises another. The recurrent desiderata core set are:

- **Validity** requires that $f(\mathbf{x}') = y'$, i.e., the counterfactual actually achieves the desired prediction. Note that the soft formulation does not guarantee it: for a given λ, the algorithm may converge without flipping the class.
- **Proximity** requires $d(\mathbf{x}, \mathbf{x}')$ to be small. The chosen distance directly affects this property: L_1 penalizes deviations linearly, encouraging sparser solutions where fewer features change substantially; while L_2 penalizes changes quadratically, favoring small changes spread across features. For mixed-type features, Gower distance is commonly used [Verma et al. 2024].
- **Actionability** requires that the changes are feasible given real-world constraints.

- **Sparsity** requires that few features are modified, which is cognitively motivated: humans naturally focus on a small number of causes rather than a complete causal chain [Miller 2019], and a recommendation to change many variables simultaneously is neither interpretable nor actionable. Generally, L_0 or L_1 (a convex approximation of L_0) are used to enforce this property.

Methods to generate CFEs can be varied, differing along three dimensions: *where* they search for x' (e.g., the full feature space, density-restricted regions, the training set itself, or a local surrogate model); *how* they navigate the space (e.g., gradient descent, stochastic search, genetic algorithm, reinforcement learning); and *what information* they use beyond querying f as an oracle (only the classifier, aggregate training statistics, individual data points, or model internals).

Optimization-based methods such as DiCE [Mothilal et al. 2020] frame the search as a weighted loss minimization over the feature space. While highly flexible, tuning the multiple competing weights in the loss function has computational cost and can affect convergence. Heuristic approaches search the decision boundary through local perturbation; Growing Spheres (GS) [Laugel et al. 2018], for instance, expands spheres around the original input independently of the training distribution, finding minimal changes efficiently but without awareness of plausibility. Instance-based methods retrieve or stay close to existing training points, guaranteeing plausibility by construction; NICE, for example, retrieves the nearest training point with the desired class, at the cost of being limited to what the training data contains. Data-manifold methods restrict the search to high-density regions of the training distribution; FACE [Poyiadzi et al. 2020], for example, follows a graph-based representation of feasible paths, producing plausible counterfactuals but at potentially high computational cost and with reduced flexibility under hard mutability constraints. Surrogate-based methods fit a local interpretable model around the input and extract the counterfactual from its structure; LORE follows this approach, requiring access to model internals or training data to fit the surrogate. Since no single method simultaneously satisfies all desiderata [Moreira et al. 2025], generation method choice involves explicit trade-offs between desiderata, computational cost, and coverage.

3. Related Work

3.1. Predictive Analytics in Education

[Suaza-Medina et al. 2024] applied SHAP to a model trained on Colombia's *Saber 11* exam and found that the socioeconomic index, gender, region, and localization were the dominant predictors in lagging regions. These findings are corroborated by [Dutra et al. 2023], whose review identifies income, age, sex, race, and parental education as the most recurrent predictors of Enem performance, and by [Melo et al. 2021], who finds scholarship rates, income, race, and maternal education explain most of the score variation at the municipal level. While these studies successfully identify structural inequalities and predictive patterns, the most influential variables are predominantly immutable or only actionable in the long term. [Silva Filho et al. 2023] incorporated causal reasoning to mitigate bias when analyzing Brazilian secondary education, while [de Jesus et al. 2025] combined SHAP with Granger causality and found that restrooms and multimedia equipment temporally precede preschool enrollment growth. Although those works identify global factors, they provide no instance-level guidance towards improvement.

3.2. Counterfactual Explanations in Education

[Tsiakmaki and Ragos 2021] presented a case study of CFEs for student academic performance, describing how a student could alter the model's prediction. [Afrin et al. 2023] applied DiCE to suggest changes in social media usage behavior that could flip a student's prediction from fail to pass, validating the recommendations through a user study with 18 educators, who found the explanations useful but raised feasibility concerns. Despite these advances, the systematic review by [Buñay-Guisñan et al. 2026], covering 29 primary studies on CFEs in education between 2017 and 2024, identifies a central weakness: most methods aim to change the predicted class at any cost, without considering actionability, fairness, or trustworthiness, a gap addressed by only two of the reviewed studies. Another work by the same authors [Buñay-Guisñan et al. 2025] further notes that most CFE research in education focuses on individual-level scenarios, even though many students or schools share characteristics that could benefit from a common intervention.

3.3. Benchmarking CFE Methods

[Moreira et al. 2025] tested DiCE, WachterCF, Prototype, and GrowingSpheresCF on white-, grey-, and black-box models, varying the underlying model across 25 datasets, and found that model architecture has little impact on CFE generation, that proximity-only algorithms (such as WachterCF) generate CFEs that are mathematically close but meaningless, and that quantitative metrics alone, without a plausibility check, lead to biased conclusions. Thus, they also recommend qualitative "counterfactual inspection". In the educational domain specifically, [Cavus and Kuzilek 2024] conducted the first benchmark of CFE methods (MOC, NICE, and WhatIf) on the OULAD dataset for student success prediction, finding that NICE optimized for sparsity provided the best overall results, while MOC was the only method to occasionally produce invalid CFEs.

3.4. SHAP vs CFE

[Mothilal et al. 2021] provided a formal demonstration that feature attribution and CFEs frequently disagree: the top-k features ranked by LIME or SHAP are often neither *sufficient* (a highly-ranked feature may not flip the decision alone) nor *necessary* (the decision may flip through features that ranked low in importance). [Kumar et al. 2020] demonstrate that Shapley-value-based feature importance carries unresolved mathematical ambiguities. Specifically, due to feature correlation, SHAP frequently assigns high importance scores to proxy variables that have no direct interventional effect on the model's output. Even within attribution methods themselves, agreement is not guaranteed. [Swamy et al. 2023] validated LIME and SHAP explanations for student success prediction through 26 interviews with university educators, finding that the methods significantly disagree with each other on which features matter the most. Moreover, the educators themselves did not agree on which explanations were most trustworthy.

4. Methodology

The proposed pipeline consists of: (i) data integration and preprocessing, (ii) performance tier construction, (iii) model training and selection, (iv) counterfactual generation under different mutability constraint scenarios, (v) evaluation and subgroup analysis, including comparison with SHAP, and (vi) diverse counterfactual generation and metrics analysis.

Table 1. Dataset distribution and actionability by subgroup.

Category	Train		Test		Action.
	n	%	n	%	%
Region					
North	1,139	9.2	280	9.1	26.4
Northeast	3,480	28.2	913	29.5	37.2
Southeast	4,740	38.4	1,172	37.9	67.0
South	1,976	16.0	480	15.5	55.2
Central-West	1,023	8.3	245	7.9	52.4
Localization					
Urban	11,476	92.9	2,896	93.7	51.2
Rural	882	7.1	194	6.3	19.6

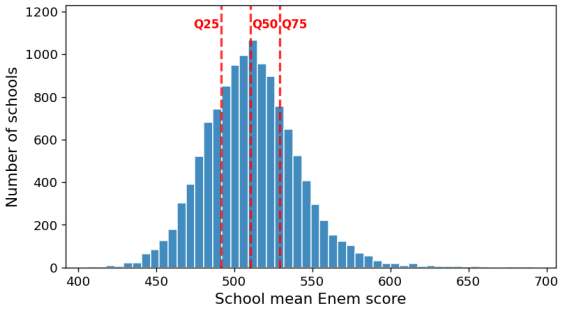


Figure 1. Training distribution. Thresholds of Q25: 491.7, Q50: 510.4, and Q75: 529.2.

4.1. Data Sources and Integration

We combine two main data sources: Enem 2024 and Brazilian School Census 2024[1]. The school-level performance is computed as the average Enem score per school, considering only students present in all exams with a valid essay score. Inactive schools or with less than 10 valid participants were excluded. The Census provides structural and administrative information. We also incorporate four school-level indicators published by INEP: Teacher Adequacy Index (`AFD`), Average Daily Hours (`HAD`), Age-Grade Distortion Rate (`TDI`), and Teacher Regularity Index (`IRD`), capturing teacher qualification, instructional time, student progression, and staff stability, respectively.

We restrict our analysis to public state schools, ensuring consistency in governance structure and making the intervention scenario realistic. Private or federal schools operate under different accountability frameworks, which could generate infeasible CFEs. The dataset split into training (80%) and test (20%) resulted in 12,358 training schools and 3,090 test schools. Preprocessing was performed with all transformations fit exclusively on the training set, and imputing missing values considered median for numerical features and mode for categorical ones. Hyperparameter tuning for all 5 classifiers (Random Forest, Logistic Regression, XGBoost, LightGBM, and MLP) used 5-fold cross-validation.

School performance is modeled as a 4-class classification problem, with the target (`TIER`) constructed using quantile-based discretization of the school's average Enem score. This yields four balanced classes, using quartile thresholds of 491.7, 510.4, and 529.2 points. The interquartile range spans 38 points, meaning adjacent tiers are not separated by pedagogically obvious boundaries. Such fixed thresholds would produce severely imbalanced classes, hence difficulting the counterfactual generation. Quartile-based tiers yield balanced classes at the cost of relative rather than absolute interpretability: the goal is to move a school from its current quartile to the next, not to reach a specific absolute score. We do not apply synthetic oversampling to further balance classes, as doing so could compromise counterfactual plausibility. Finally, the geographic and localization

[1]Enem, School Census and INEP (Inst. Nacional de Estudos e Pesq. Educacionais Anísio Teixeira) indicators dataset are available at `https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/`. Additional socioeconomical information was retrieved from `https://www.ipeadata.gov.br/` and `https://sidra.ibge.gov.br/home/pmc/brasil`. Detailed descriptions of all dataset attributes are provided in `https://github.com/anareboucas/cfe`.

distributions are consistent between training and test sets, as shown in Table 1.

4.2. Feature Mutability Framework

Features are organized into three mutability levels, as follows:

- **Immutable** features (10) cannot be changed through direct intervention. These include geographic attributes, administrative type, enrollment count, whether the school serves indigenous communities, and contextual socioeconomic indicators.
- **Structural** features (49) require medium-to-long term investment or systemic policy changes to modify. Examples include the presence of a science laboratory and if the school operates in its own building. Also, indicators that are not directly actionable, but can change by acting towards it, such as age-grade distortion rate (`TDI`), which reflect longer-term policy outcomes, are considered in this category.
- **Operational** features (60) are directly actionable at the school administration level in the short term, such as the number of computing devices for students.

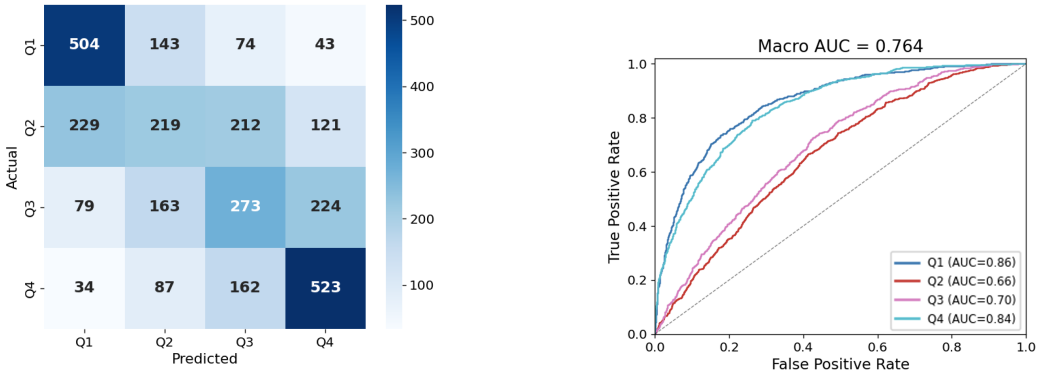
4.3. Counterfactual Generation

CFEs are generated for test set instances, targeting the next performance tier for each school. Schools already in the top tier (predicted Q4) are excluded, as no upward target exists. DiCE is used as our primary method, since it natively supports feature mutability constraints. Although it supports diversity, we did not consider it to keep comparability across methods. DiCE's generated CFE goes through the following post-processing: (i) immutable features are restored to the query instance's original values, correcting any drift introduced during the genetic search; (ii) binary/count feature pair constraints are made consistent (e.g., if `QT_DESKTOP_ALUNO` is above zero, `IN_DESKTOP_ALUNO` is set to 1); (iii) the CFE is re-evaluated by the classifier and only counted as *actionable* if it still predicts the desired tier. This distinguishes *validity rate* (before correction) from *actionability rate* (after correction). GS is evaluated as a data-agnostic alternative, expanding spheres of initial radius 0.1 around the original instance over up to 20 iterations until reaching the desired class. Candidate points are generated by adding Gaussian noise scaled by the per-feature standard deviation for continuous features, and, in the case of categorical features, by sampling uniformly from their domain. Immutable features are held fixed. After a candidate is found, the same post-processing pipeline (immutable restoration, binary/count pair synchronization, re-validation) is applied. For FACE, a 10-NN graph is built over training instances. Counterfactual candidates are restricted to instances from the same region and localization as the query to ensure that immutable geographic plausibility. The same post-processing pipeline is then applied to the returned instance.

4.4. CFE Evaluation Metrics

Metrics are computed per instance and aggregated by mutability scenario and subgroup:

- **Validity rate**: proportion of instances with at least one CFE predicted for the target class before post-processing corrections, measuring raw algorithmic success.
- **Actionability rate**: proportion of instances with at least one valid CFE that remains valid after the post-processing steps. This is the primary reported metric.
- **Proximity**: Gower distance between the original instance and its CFE.
- **Sparsity**: number of features changed, excluding redundancies (e.g., changing `QT_DESKTOP_ALUNO` and `IN_DESKTOP_ALUNO` counts as one intervention).
- **Execution time**: total time (sec) required to generate CFEs for all query instances.



(a) Confusion Matrix **(b) ROC Curve**

Figure 2. Performance evaluation of the XGBoost classifier.

Table 2. Example CFE for a Q2 school (State High School Raymundo Martins Vianna, Belém - PA) targeting Q3 (operational scenario).

Feature	Orig.	CFE	Feature	Orig.	CFE
Support staff			**Pedagogical equipment**		
Psychologist Count (QT)	0	1	Multimedia Equip. Count (QT)	1	6
Pedagogy Staff Count (QT)	5	9	Printer (IN)	0	1
Librarian Count (QT)	0	1	Copier (IN)	0	1
Monitor Count (QT)	0	1	**Teachers**		
Accessibility			Teacher Count (QT)	34	44
Visual Access. Signal (IN)	1	0			

5. Results and Discussion

5.1. Dataset and Models Performance

XGBoost achieves the highest macro F1 on the test set (0.476), followed closely by Random Forest (0.474), LightGBM (0.471), Logistic Regression (0.463), and MLP (0.448), and is therefore selected as the primary model for CFE generation. Although we select the model with the best metric, we noticed that higher predictive performance does not necessarily translate into better CFE quality. For instance, a model can achieve high accuracy by exploiting class concentration, correctly classifying schools in dense, homogeneous regions of the feature space, precisely where small operational changes are least likely to cross the decision boundary. Conversely, where the decision boundary is poorly separated, as expected given the narrow interquartile range, adjacent-tier schools differ only marginally, making a CFE less reliable as a meaningful recommendation. This reflects how the shape of the learned decision boundary, not just its accuracy, determines how accessible it is through minimal feature changes, with direct implications for model selection in XAI pipelines, which should incorporate downstream CFE quality when the goal is actionable explanation. Figure 2 confirms that extreme-tier schools are easier to separate than schools near the narrow interquartile range, as expected.

5.2. Comparison with SHAP

The alignment between SHAP importance and CFE frequency is quantified using the Spearman rank correlation coefficient, which is $\rho = 0.16$, under the operational scenario, suggesting a divergence between predictive relevance and actionable change. The features

Table 3. SHAP importance (normalized among mutable features) vs. CFE frequency (left). Frequency of modified subcategories (right).

SHAP Top 10 Features	%	CFE Top 10 Features	%	Subcategory (# items)	% schools	% changes
Operational				*Operational*		
Multimedia eq. count	11.42	Teacher count	96.5	Pedagogical equip. (9)	99.4	21.0
Student desktop count	9.96	Student desktop count	91.7	Support staff (8)	98.4	23.7
Student tablet count	6.88	Multimedia eq. count	86.2	Student equipment (3)	97.9	15.2
Teacher count	6.60	Student laptop count	85.6	Teachers (1)	96.4	6.2
Management staff count	6.24	Pedagogy staff count	81.9	Ped. materials (10)	92.6	16.4
Student laptop count	5.60	**Audio equip. count**	81.4	Connectivity (7)	70.6	7.3
Science ped. materials	4.15	Management staff count	69.8	Governance (3)	60.7	5.1
Musical ped. materials	4.04	**Librarian staff count**	62.4	Accessibility (4)	39.1	3.1
Pedagogy staff count	3.66	Student tablet count	58.8	Services and env. (1)	30.7	2.0
Secretary staff count	3.53	**Coordinator count**	54.8			

that most influence the model are age-grade distortion (TDI_MED), geographic region (CO_REGIAO), per capita GDP (PIB_PER_CAPITA), teacher adequacy (AFD), and state (CO_UF); whereas the features that appear most frequently in CFEs have lower SHAP importance globally, but are locally sufficient to cross the decision boundary for individual schools. This reveals a fundamental gap: a policymaker reading a SHAP plot might conclude that school improvement depends on reducing poverty or changing geography, neither feasible nor actionable. The CFE output instead provides concrete, school-specific recommendations (Table 2), such as increasing support staff, teachers from 34 to 44, and pedagogical equipment. This is a step further in understanding: an actionable path that can be budgeted and implemented in the current school year. However, even with the mutability level restriction and careful feature selection, some generated CFEs may still suggest spurious changes, such as recommending a reduction in accessibility or equipment availability. This may reflect indifference toward that attribute rather than an actual detrimental effect, or it may capture a spurious effect. Also, CFEs identify features whose change is associated with a different model prediction, not causal relationships: they are intervention hypotheses, not guarantees. The SHAP top features among operational levers are dominated by equipment counts, while CFEs concentrate on staffing and equipment (Table 3); in the structural scenario, school indicators and infrastructure rank first.

5.3. Effect of Mutability Constraints

Table 4 shows that actionability rates decrease as constraints become stricter, regardless of the generation method. Although we have 60 mutable features on the operational scenario, 14 are binary indicator flags (IN) redundant with their quantitative (QT) counterparts. Using DiCE, for approximately 50% of schools, the operational levers are insufficient to move the school to the next performance tier, which reflects a structural challenge. Structural investment (e.g. in physical infrastructure and faculty qualifications) is a prerequisite for these schools before operational changes can be effective. The absence of a valid CFE is therefore itself an informative result, distinguishing schools where targeted operational intervention is viable from those where systemic support is needed first. The relationship between constraint strictness and execution time brings scalability concerns. DiCE is roughly 8 times slower under operational constraints than under no restrictions. This counterintuitive result, a smaller search space taking longer, is due to the nature of genetic search, where with fewer mutable features, each step has less room to maneuver, requiring more generations to cross the decision boundary. For many schools, no such

Table 4. CFE generation methods and scenarios comparison.

Method	Scenario	Mutable features	Validity/ Actionability (%)	Proximity (Gower)	Sparsity (features)	Time (s)
DiCE	Operational	46	48.1 / 47.8	0.088	15.4	14087.3
DiCE	Structural	95	93.9 / 92.7	0.153	31.9	3456.4
DiCE	Unrestricted	103	99.6 / 98.3	0.164	35.2	1849.2
GS	Operational	46	88.6 / 87.9	0.021	5.0	1108.7
GS	Structural	95	100.0 / 88.5	0.019	5.9	1152.2
FACE	Operational	46	14.2 / 14.2	0.056	12.3	78.2
FACE	Structural	95	66.2 / 65.3	0.121	28.6	70.9

direction exists and the generation budget is exhausted before convergence.

Proximity and sparsity both increase as constraints are relaxed: DiCE changes more features (35.2 vs. 15.4) and accumulates larger Gower distance (0.164 vs. 0.088). The growing number of changed features drives the aggregate proximity up, even when the per-feature cost can actually drop as constraints relax, which would be expected by the sparsity-proximity trade-off. The high actionability of GS should not be interpreted as high feasibility: the method generates synthetic points by random perturbation in feature space, which may not correspond to realistic school configurations, unlike FACE which returns actual training instances.

5.4. CFE Methods Comparison

The methods show trade-offs consistent with their design (Table 4). GS achieves the highest actionability under operational constraints (87.9%), while DiCE wins under structural and unrestricted scenarios (92.7-98.3%). DiCE requires higher computational cost under tight constraints, while GS, under operational constraints, achieves both high validity and high actionability and a lower time spent. Increasing features raises feature-selection cost, but structural's higher validity lowers generation cost by avoiding failed searches, consistent with GS's similar execution time. FACE shows the lowest validity and actionability. Because it returns actual training instances, it guarantees plausibility, but this comes at the cost of requiring a real school with the desired class, a reachable density-weighted path from the query, and matching categorical immutable features. This combination is rarely satisfied, which explains the low validity. The fast execution is a consequence of the graph being pre-computed once, requiring only a shortest-path search per query.

5.5. Subgroup Analysis

Schools in lower-resource regions (N and NE) appear more frequently among those without valid CFEs under the operational scenario. This is expected, since schools starting from a lower baseline often require larger cumulative changes across changing dimensions to reach the next tier, which may exceed what is feasible. Rural schools demonstrate a similar pattern, with lower actionability rates than urban schools, due to the great resource gap that characterizes them in Brazil (fewer computing devices, teachers, and internet connectivity). Thus, schools that need the most intervention are also the least likely to receive an operational counterfactual, a finding with direct implications for equitable policy design. The same reasoning applies to the tier, where schools starting in Q1 are the hardest to generate valid CFEs under operational constraints (Q1: 30.1%, Q2: 61.6%, Q3: 59.9%). These patterns, shown in Table 1, reinforce that a uniform intervention policy would systematically fail the schools that need support most. CFEs, precisely because

they are instance-level, surface this heterogeneity explicitly: the absence of a valid CFE for a school is itself an informative result, signaling that structural or systemic investment is a prerequisite before operational changes can produce tier-level improvement.

6. Conclusion

We investigate CFEs as a tool to move from prediction to actionable insight, focusing on Brazilian public state schools. Our results show that the features that better explain the model's decision are not the same whose change would improve school performance, highlighting a fundamental gap between predictive importance and actionable change, i.e., SHAP-ranked features tend to be socioeconomic and immutable, while CFE-recommended features can be operational and actionable. Incorporating real-world restrictions significantly reduces the proportion of valid counterfactuals, since not every school has a clear operational path to the next performance tier without prior structural investment. Rather than a limitation, this is itself a valuable policy signal, distinguishing schools where short-term action suffices from those that need broader, longer-term support. Such instance-level granularity also makes CFEs preferable to global recommendations, since actionability varies meaningfully across subgroups (region, urban/rural, initial performance). The generation method also shapes the result, while no single metric captures which one is the best overall; the right choice depends on feasibility trade-offs selected by school administrators and policymakers. Making these trade-offs explicit across every level of modeling decision is one of the central contributions of this work.

References

Afrin, F., Hamilton, M., and Thevathyan, C. (2023). Exploring counterfactual explanations for predicting student success. In *ICCS 2023*, pages 413–420. Springer.

Bayrak, B. and Bach, K. (2024). Evaluation of instance-based explanations: an in-depth analysis of counterfactual evaluation metrics, challenges, and the ceval toolkit. *IEEE Access*, 12:137683–137695.

Buñay-Guisñan et al. (2025). Group counterfactual explanations: A use case to support students at risk of dropping out in online education. *Electronics*, 15(1):51.

Buñay-Guisñan, P., Lara, J., and Romero, C. (2026). Counterfactual explanations in education: A systematic review. *WIREs Data Mining and Knowledge Discovery*, 16(1):e70060.

Cavus, M. and Kuzilek, J. (2024). The actionable explanations for student success prediction models: A benchmark study on the quality of counterfactual methods. In *EDM*.

de Jesus, A. C. et al. (2025). Understanding public school enrollment trends in the northeast of brazil: Unraveling features through ai related to the learning process for low-income students. In *AIED 2025*, pages 454–461. Springer.

Dutra, J. F., Júnior, J. B. F., and de Souza Fernandes, D. Y. (2023). Fatores que podem interferir no desempenho de estudantes no enem: uma revisão sistemática da literatura. *Revista Brasileira de Informática na Educação*, 31:323–351.

Guidotti, R. (2024). Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824.

Karimi, A.-H., Schölkopf, B., and Valera, I. (2021). Algorithmic recourse: from counterfactual explanations to interventions. In *ACM FAccT 2021*, pages 353–362.

Khosravi, H. et al. (2022). Explainable artificial intelligence in education. *Computers and education: artificial intelligence*, 3:100074.

Kumar, I. E. et al. (2020). Problems with shapley-value-based explanations as feature importance measures. In *ICML 2020*, pages 5491–5500. PMLR.

Laugel, T. et al. (2018). Comparison-based inverse classification for interpretability in machine learning. In *IPMU 2018*, pages 100–111. Springer.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *NeurIPS 2017*.

Melo, R. O. et al. (2021). Impacto das variáveis socioeconômicas no desempenho do enem: uma análise espacial e sociológica. *Rev. de Admin. Pública*, 55(6):1271–1294.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Moreira, C. et al. (2025). Benchmarking instance-centric counterfactual algorithms for xai: from white box to black box. *ACM Computing Surveys*, 57(6):1–37.

Mothilal, R. et al. (2021). Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *AAAI/ACM AIES 2021*, pages 652–663.

Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *ACM FAccT 2020*, pages 607–617.

Poyiadzi, R. et al. (2020). Face: feasible and actionable counterfactual explanations. In *AAAI/ACM AIES*, pages 344–350.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *ACM SIGKDD 2016*, pages 1135–1144.

Silva Filho, R. L. C., Brito, K., and Adeodato, P. J. L. (2023). Leveraging causal reasoning in educational data mining: an analysis of brazilian secondary education. *Applied Sciences*, 13(8):5198.

Suaza-Medina, M., Peñabaena-Niebles, R., and Jubiz-Diaz, M. (2024). A model for predicting academic performance on standardised tests for lagging regions based on machine learning and shapley additive explanations. *Scientific Reports*, 14(1):25306.

Swamy, V., Du, S., Marras, M., and Kaser, T. (2023). Trusting the explainers: teacher validation of explainable artificial intelligence for course design. In *LAK 2023*.

Tsiakmaki, M. and Ragos, O. (2021). A case study of interpretable counterfactual explanations for the task of predicting student academic performance. In *CSCC 2021*.

Ustun, B., Spangher, A., and Liu, Y. (2019). Actionable recourse in linear classification. In *ACM FAccT 2019*, pages 10–19.

Verma, S. et al. (2024). Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, 56(12):1–42.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.