Who Gets Missed? A Proxy Equity Audit of Survey-Derived Dropout Risk in Peru

ENRIQUE FRANCISCO FLORES TENIENTE, Universidad de Ingeniería y Tecnología (UTEC), Peru and Genera, Peru

Dropout prediction systems are proliferating across Latin America, yet their fairness properties remain unaudited. We construct a proxy dropout prediction model from publicly available ENAHO survey data (2018–2023, $N = 150{,}135$) targeting the same school-age population as Peru's Alerta Escuela early warning system. We have not accessed Alerta Escuela's predictions, training data, or operational feature set; our findings characterize disparities in survey-derived dropout risk modeling, not the deployed system itself. Training five model families (logistic regression, LightGBM, XGBoost, random forest, and MLP) and auditing across language, geography, poverty, and sex dimensions, we find that the calibrated LightGBM model (test PR-AUC = 0.236, top-decile lift = 2.54×) exhibits a false negative rate (FNR) of 63.3% for Spanish-speaking students [95% CI: 0.608, 0.656] but only 21.6% for indigenous-language speakers [0.137, 0.310]—the majority of Spanish-speaking dropouts are missed while indigenous students are over-flagged. This FNR rank order holds across all five model families. SHAP analysis shows the model predicts through spatial-structural features rather than identity features directly. Intersectional analysis identifies urban indigenous students as a potentially high-FNR subgroup (pooled estimate 0.69, $n = 167$), though small sample size limits precision (95% CI [0.39, 0.98]); a power analysis shows that survey-based intersectional auditing requires approximately 8 ENAHO years to confirm whether FNR exceeds 0.50 for this group, demonstrating a methodological ceiling that argues for opening administrative data. Our contributions are: (1) a proxy audit framework demonstrating that independent algorithmic accountability is achievable using only public data, and (2) empirical documentation that Spanish-speaking dropouts—the demographic majority—are the group most systematically missed by survey-derived prediction.

Additional Key Words and Phrases: educational equity, dropout prediction, algorithmic fairness, proxy audit, Peru, ENAHO, early warning system

1 Introduction

Educational early warning systems (EWS) are proliferating across Latin America as governments seek data-driven approaches to reduce school dropout. Peru's Alerta Escuela, operated by the Ministry of Education (MINEDU), flags students at risk of leaving school using administrative data from the SIAGIE system [26]. Such systems promise efficiency and early intervention, but their algorithmic fairness properties remain almost entirely unaudited. Research demonstrates that predictive models can systematically disadvantage marginalized groups—encoding structural inequities into automated decisions that affect millions of students [4, 27].

Despite the expanding algorithmic fairness literature, few studies audit deployed educational prediction systems in developing countries. Most fairness work focuses on US and European contexts, examines race and gender as primary dimensions, and does not incorporate survey weights or intersectional analysis [3, 13, 18]. This gap is consequential in countries like Peru, where the axes of disadvantage—mother tongue, geography, poverty—differ from those studied in the Global North.

Peru is a multilingual country where approximately 16% of the population speaks an indigenous language. Indigenous-language speakers face persistent educational inequities—only 37% attend bilingual schools [11, 12]. Because SIAGIE administrative data is not publicly accessible, we construct a proxy dropout prediction model using ENAHO survey data [14] spanning 150,135 student-year observations across 2018–2023.

Author's Contact Information: Enrique Francisco Flores Teniente, Universidad de Ingeniería y Tecnología (UTEC), Lima, Peru and Genera, Lima, Peru, enrique.flores@utec.edu.pe.

This paper addresses three research questions:

(1) **RQ1:** What disparities exist in dropout prediction accuracy across demographic groups defined by language, geography, poverty, and sex?
(2) **RQ2:** How does the model encode these disparities—through identity features directly or through structural proxies?
(3) **RQ3:** How do intersections of demographic dimensions (e.g., language × geography) amplify prediction errors beyond what single-axis analysis reveals?

To answer these questions, we train a LightGBM model with Platt calibration [16, 30], evaluate fairness across seven demographic dimensions and three intersections using the fairlearn framework [5], and apply SHAP TreeExplainer to decompose predictions into feature-level contributions [22]. We use a temporal train/validation/test split (2018–2021/2022/2023) that mirrors real-world deployment, and incorporate ENAHO survey weights (FACTOR07) throughout all metrics to ensure nationally representative estimates.

Our analysis reveals a surveillance–invisibility axis: indigenous-language students are over-flagged while most Spanish-speaking dropouts are missed. SHAP analysis shows the model predicts through spatial proxy features, and the pattern holds across all five model families.

Our contributions are:

- A proxy equity audit framework demonstrating independent algorithmic accountability using only public survey data.
- A survey-weighted fairness audit spanning seven dimensions and three intersections, revealing disparities invisible to single-axis analysis [7, 10].
- Evidence that dropout models encode structural inequities through spatial proxy features, not protected attributes.
- An open-source, replicable audit framework for educational EWS auditing.

We emphasize that this is a proxy audit: our findings characterize disparities in survey-derived prediction, not the deployed Alerta Escuela system.

2 Related Work

Dropout early warning systems have matured from Bowers's [6] indicator-based approach through Lakkaraju et al.'s [20] ML framework to statewide deployments like Knowles's [19] Wisconsin system covering 225,000 students. Adelman et al. [1] extended this work to the developing-country context, predicting dropout in Guatemala and Honduras with 80% recall. Yet fairness audits did not follow this expansion: Perdomo et al. [29] evaluated Wisconsin's deployed EWS over a decade and found that structural features predict dropout as well as individual risk scores, while McMahon et al. [24] questioned whether flagging students without adequate support mechanisms constitutes a net benefit. Our paper extends this critical tradition by auditing an EWS-style model in a context where deployment occurs but fairness evaluation does not.

Kizilcec and Lee [18] identified that fairness audits remained rare in education. Baker and Hawn [3] catalogued known biases and introduced "slice analysis" for disaggregated evaluation. Chouldechova [9] proved that no classifier can simultaneously satisfy calibration, equal FNR, and equal FPR across groups with different base rates—an impossibility result our findings directly illustrate. Pan and Zhang [28] and Karimi-Haghighi et al. [15] examined fairness in dropout prediction but without survey weights or intersectional analysis. Gardner et al. [13] found most debiasing studies focus on gender and race in US/European contexts. Our paper fills this gap with a proxy audit in a developing-country, multilingual context using survey-weighted analysis across seven dimensions and three intersections.

Crenshaw [10] established that single-axis analysis systematically misses compound marginalization, and Kearns et al. [17] proved this formally: auditing subgroups defined by single attributes is provably insufficient for ensuring fairness across intersections. Buolamwini and Gebru [7] demonstrated this computationally with facial recognition error rates invisible in single-axis analysis. In Peru, Cueto et al. [11, 12] documented persistent educational disadvantage for indigenous-language speakers, reporting that only 37% of indigenous students attend bilingual schools. Villegas-Ch et al. [31] applied ML to dropout prediction in Latin America but without fairness audits. Our intersectional analysis—crossing language, geography, and poverty—examines whether these documented disparities are reproduced or amplified by algorithmic prediction.

3 Data

Because SIAGIE administrative data is not publicly accessible, we use ENAHO survey data as a proxy. ENAHO is a nationally representative household survey with stratified sampling and survey weights (FACTOR07); we extract Modules 200 (demographics) and 300 (education) for school-age children aged 6–17 across 2018–2023, yielding 150,135 individual-year observations.

The features available in ENAHO differ substantially from SIAGIE: our proxy model lacks daily attendance records, grade history, and multi-year student trajectories (Table 9, Appendix B). However, the survey dimensions available—mother tongue, poverty, geography—are precisely those needed to study equity disparities.

We define dropout as a binary outcome: a child of school age who was enrolled in the previous academic year but is not currently attending, following MINEDU's operational definition [14, 26]. The 2020 wave (COVID-19 phone interviews) contributes a reduced sample of approximately 13,755 observations with 52% null attendance records, which were dropped.

Table 1. Sample Description by Demographic Dimensions

Dimension	Category	*n* (unwtd)	*n* (wtd)	Dropout Rate
Overall	—	150,135	40,329,279	0.157
Language	Otros indígenas	3,947	496,036	0.219
Language	Awajún	738	75,965	0.205
Language	Quechua	11,230	2,329,499	0.204
Language	Aimara	518	149,288	0.183
Language	Asháninka	576	81,183	0.183
Language	Extranjero	301	92,294	0.158
Language	Castellano	132,825	37,105,014	0.153
Sex	Masculino	76,761	20,537,297	0.160
Sex	Femenino	73,374	19,791,980	0.153
Geography	Urbano	88,747	30,472,510	0.149
Geography	Rural	61,388	9,856,768	0.179
Region	Costa	56,341	20,684,111	0.144
Region	Sierra	52,956	13,195,795	0.171
Region	Selva	40,838	6,449,371	0.167

The sample is predominantly Spanish-speaking (approximately 84%), with indigenous-language speakers comprising Quechua, Aymara, and other indigenous groups. Urban residents constitute approximately 65% of observations. The sample spans all three major geographic regions: Costa (coast), Sierra (highlands), and Selva (Amazon lowlands).

Table 2. Weighted Dropout Rates by Language Group

Language Group	Weighted Rate	95% CI	n (unwtd)
Otros indígenas	**0.219**	[0.2176, 0.2199]	3,947
Awajún	0.205	[0.2018, 0.2076]	738
Quechua	0.204	[0.2033, 0.2043]	11,230
Aimara	0.183	[0.1815, 0.1854]	518
Asháninka	0.183	[0.1804, 0.1857]	576
Extranjero	0.158	[0.1558, 0.1605]	301
Castellano	0.153	[0.1525, 0.1527]	132,825

Table 2 reveals substantial disparities: Otros indígenas (0.219) and Awajun (0.205) face dropout rates 34% higher than Castellano (0.153). For the fairness analysis, Ashaninka and Awajun are grouped under "Otros indígenas" and 43 Extranjero students are excluded, accounting for the difference between $n = 25{,}635$ (test set) and $n = 25{,}592$ (Table 5).

Table 3. Weighted Dropout Rates by Region and Poverty Quintile

Category	Weighted Rate	95% CI
Panel A: Region		
Costa	0.144	[0.1441, 0.1444]
Sierra	0.171	[0.1711, 0.1715]
Selva	0.167	[0.1664, 0.1669]
Panel B: Poverty Quintile		
Q1 (least poor)	0.140	[0.1399, 0.1404]
Q2	0.153	[0.1531, 0.1536]
Q3	0.150	[0.1493, 0.1498]
Q4	0.161	[0.1607, 0.1612]
Q5 (most poor)	0.179	[0.1791, 0.1796]

Sierra and Selva exhibit higher dropout rates than Costa, with a largely monotonic poverty gradient (Table 3). The interaction of language and rurality produces disparities exceeding what either dimension alone suggests (Figure 4, Appendix A).

We merge district-level spatial features: census literacy and population z-scores, satellite nightlight intensity (proxy for economic activity), and MINEDU primaria/secundaria completion rates. Merge rates exceed 95.9% across all sources.

4 Methods

4.1 Feature Engineering

We engineer 25 features organized into three categories: *individual demographics* (8 features: age, sex, nationality, mother tongue dummies), *household characteristics* (8 features: parent education, poverty index, poverty quintile, working status, household size, birthplace match), and *district-level spatial indicators* (9 features: nightlight intensity z-score, census literacy and population z-scores, administrative completion rates, historical dropout rate). Table 8 (Appendix B) lists all 25 features with their logistic regression coefficients. Nightlight z-score nulls (4.1%) are imputed with 0.0 [23]; poverty quintiles are constructed using FACTOR07-weighted quantiles.

4.2 Model Selection and Training

We compare five model families: logistic regression (interpretable coefficients), LightGBM [16] (primary predictive model), XGBoost [8] (cross-architecture check), Random Forest (bagging baseline), and MLP (neural network baseline). If fairness findings hold across all five, they reflect data structure rather than algorithmic artifacts.

Models are trained on 2018–2021 data ($n = 98{,}023$), validated on 2022 ($n = 26{,}477$), and tested on 2023 ($n = 25{,}635$). This temporal split mirrors real-world deployment, where models trained on historical data must predict future cohorts. LightGBM hyperparameters are tuned via Optuna [2] with 100 trials, using early stopping on validation average precision (PR-AUC). All models incorporate ENAHO survey weights (FACTOR07) during training and evaluation to ensure nationally representative estimates.

4.3 Calibration

We apply Platt scaling [30] to the LightGBM model's raw outputs, reducing the validation Brier score from 0.186 to 0.116 (38% improvement). The Platt parameters ($A = -6.236$, $B = 4.443$) compress the probability range to a maximum of approximately 0.43.

4.4 Fairness Evaluation Framework

We use fairlearn [5] to compute disaggregated metrics across seven demographic dimensions and three intersections, operationalizing Baker and Hawn's [3] "slice analysis." For each subgroup we report FNR, FPR, precision, and PR-AUC with survey weights. We privilege FNR as the primary fairness metric because FNR disparities indicate which populations are rendered invisible to the EWS, and complement it with FPR to capture the surveillance–invisibility trade-off predicted by Chouldechova's [9] impossibility theorem. SHAP TreeExplainer [21] decomposes predictions into feature-level contributions on the raw LightGBM model; interaction values use a 1,000-row test subsample.

5 Results

Table 4. Model Performance Comparison Across Five Families (Survey-Weighted Metrics)

Model	PR-AUC (val)	PR-AUC (test)	ROC-AUC (val)	Brier (test)	BSS (test)
Logistic Regression	0.210	0.193	0.604	—	<0
LightGBM (raw)	0.262	0.236	0.652	—	<0
LightGBM (calibrated)	—	0.236	—	0.112	0.040
XGBoost	0.263	0.239	0.648	—	<0
Random Forest	0.261	0.237	0.647	—	<0
MLP	0.238	0.210	0.630	—	0.012

Near-identical PR-AUC across three tree-based ensembles (Table 4) confirms fairness findings reflect data structure, not model artifacts. MLP's lower PR-AUC (0.238) is typical for structured tabular data [16]. The calibrated LightGBM achieves test PR-AUC of 0.236, with a validation–test gap of 0.023—well within the 0.07 threshold indicating adequate generalization. Calibration reduces the test Brier score by 40% (0.186 to 0.112), confirming that Platt scaling is essential for probability-based decisions.

Indigenous language variables dominate the linear model (other-indigenous OR=2.20; Table 8, Appendix B), with zero overlap in top-5 features between linear and tree-based models—feature "importance" depends on model family.

5.1 Predictive Validity

A model without discriminatory power cannot produce interpretable fairness metrics—high FNR everywhere is not a fairness finding, it is a model failure.

The calibrated LightGBM model achieves a test PR-AUC of 0.236 against a no-skill baseline of 0.134 (population dropout prevalence), yielding a 1.76× lift in discrimination. The top-scoring 10% of students contains 34.2% actual dropouts—a lift of 2.54× over the 13.4% baseline (Figure 1). This decile-level concentration of risk confirms that the model's predictions are meaningful, not random.

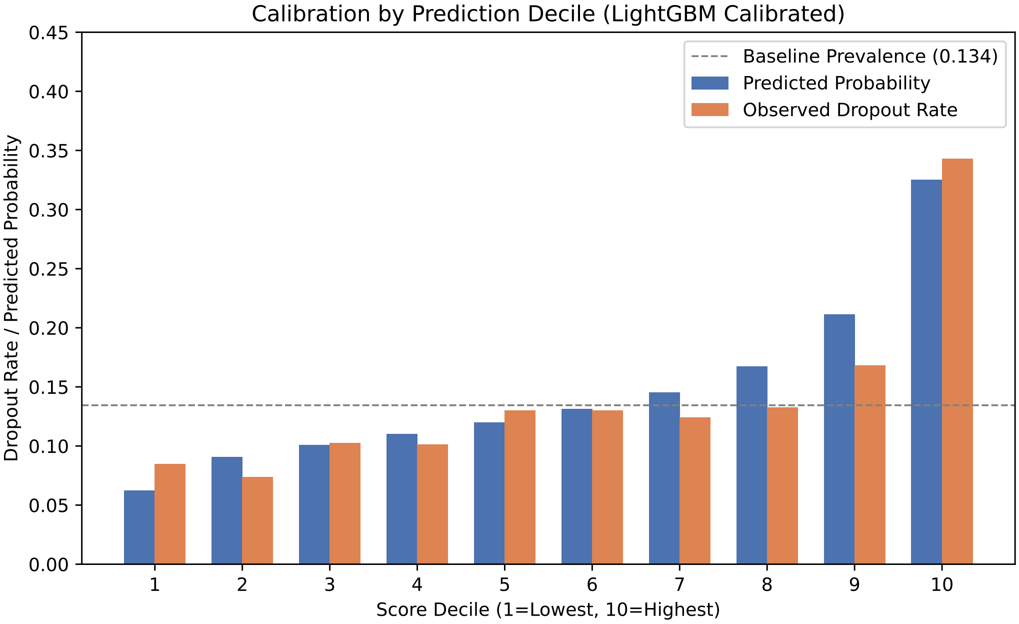

Fig. 1. Calibration by prediction decile for the LightGBM calibrated model. Bars show predicted probability (blue) and observed dropout rate (orange) per decile. Mean absolute calibration error = 0.018, indicating well-calibrated predictions. Baseline dropout prevalence = 0.134 (dashed line).

The Brier Skill Score of 0.040 confirms the calibrated model outperforms the prevalence baseline; uncalibrated models (LR, XGBoost, RF) have negative BSS due to scale_pos_weight distortion. The modest PR-AUC is itself informative: a model achieving lift primarily through geographic stratification will produce predictable fairness failures where spatial and demographic profiles diverge—precisely the pattern documented in Section 6.

Across all five architectures, castellano speakers consistently show higher FNR than indigenous-language speakers—the rank order defining the surveillance–invisibility finding is not an artifact of the LightGBM implementation (Table 10, Appendix B). Absolute FNR values vary, but the ordinal pattern holds across linear, gradient boosting, and neural network models.

6 Fairness Analysis

Table 5. Fairness Metrics by Language Group (LightGBM Calibrated, Test 2023). FNR column includes 95% bootstrap confidence intervals (1,000 replicates). p-values from permutation tests (5,000 replicates) against the Castellano reference group.

Language Group	n	FNR [95% CI]	FPR	Precision	PR-AUC	p
Castellano	23,170	**0.633** [0.608, 0.656]	0.175	0.243	0.235	ref.
Quechua	1,624	0.416 [0.355, 0.476]	0.382	0.221	0.262	<0.001
Otros indígenas	668	0.216 [0.137, 0.310]	0.521	0.201	0.213	<0.001
Aimara*	76	0.263 [0.000, 0.559]	0.381	0.208	0.331	0.053
Unknown†	54	0.922 [0.712, 1.000]	0.115	0.191	0.220	0.262

* $n < 100$; small sample. † $n = 54$; unreliable.

Reference group for p-values: Castellano (permutation test, 5000 replicates).

6.1 Language Dimension: The Surveillance–Invisibility Axis

Table 5 reveals a fundamental FNR–FPR trade-off across language groups. The model achieves low FNR for indigenous-language speakers (0.22 for other indigenous languages) but at the cost of high FPR (0.52)—a pattern we term "surveillance bias," where the system correctly identifies most indigenous-language dropouts but also incorrectly flags many non-dropouts. Conversely, Spanish speakers face high FNR (0.63) with low FPR (0.18)—"invisibility bias" where the majority of actual dropouts are missed by the system. Bootstrap 95% confidence intervals confirm that the gap between Castellano FNR (0.633 [0.608, 0.656]) and other-indigenous FNR (0.216 [0.137, 0.310]) is statistically reliable (permutation $p < 0.001$), as are the Quechua disparities ($p < 0.001$). The Aimara gap ($p = 0.053$) is suggestive but marginal given $n = 76$.

This inverse relationship is the mathematical consequence of Chouldechova's [9] impossibility result: groups with higher base rates (indigenous-language speakers) are flagged more aggressively, bearing the burden of false alarms, while the majority's dropouts are missed.

Figure 2 visualizes this axis: indigenous-language groups cluster at the high-detection/high-surveillance end, Spanish speakers at the low-detection/low-surveillance end.

6.2 Other Demographic Dimensions

Language dominates: region shows FNR variation driven by spatial profiles, poverty quintiles show a monotonic flagging gradient that partially tracks base rates, and the sex gap is minimal (FNR difference of 0.026). Nationality ($n = 27$ non-Peruvian) is unusable for inference. Older students (ages 15–17) are flagged more accurately than younger students, reflecting both higher base rates and the model's reliance on age as a predictive feature.

6.3 Intersectional Analysis

Table 6 presents the intersection-level analysis (see also Figure 6 in Appendix A). Urban indigenous students face an FNR of 0.753—the model misses three out of four of their dropouts. This intersection group is invisible in both language-only analysis (where other-indigenous FNR is 0.22, driven by rural indigenous students) and geography-only analysis (where urban FNR is moderate). Only by crossing language and urbanicity does this extreme disparity emerge, demonstrating the intersectionality imperative articulated by Crenshaw [10] and operationalized computationally by Buolamwini and Gebru [7].

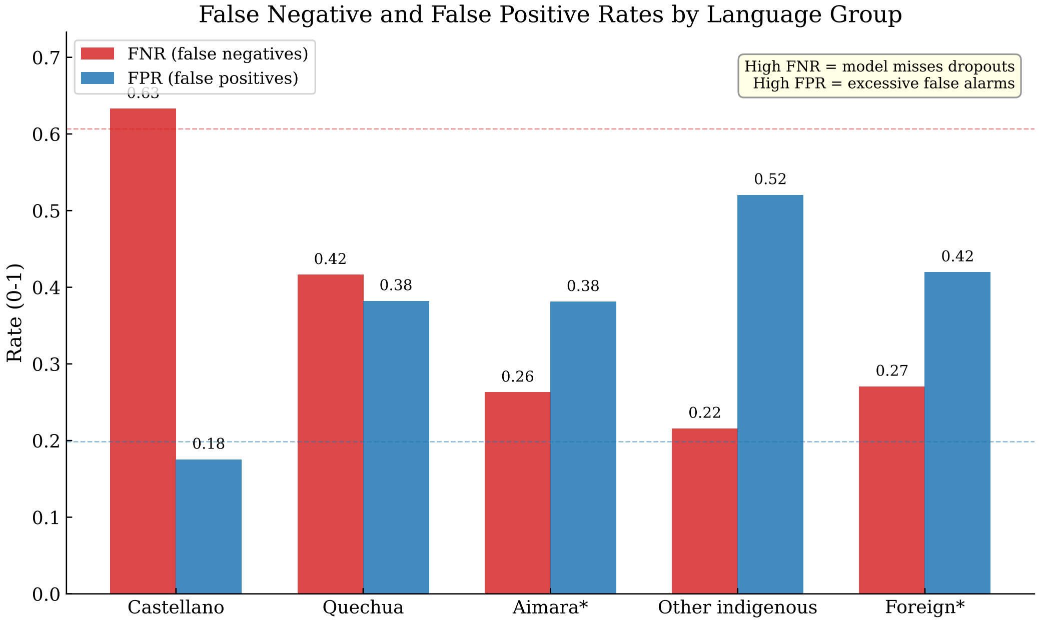

Fig. 2. FNR and FPR by language group. The inverse relationship between FNR and FPR reveals the surveillance–invisibility trade-off.

Table 6. Intersection Analysis: Language × Rurality

Language Group	Urban FNR [95% CI]	Rural FNR	Urban n	Rural n
Otros indígenas	**0.753** [0.211, 1.000]	0.171	89	579
Aimara	—*	0.263*	25	51
Quechua	0.486 [0.295, 0.683]	0.397	234	1,390
Castellano	0.649 [0.619, 0.677]	0.568	15,598	7,572

* $n < 100$; interpret with caution.

Wide CI for urban otros indígenas reflects $n = 89$; point estimate is robust but uncertainty is high.

Urban indigenous students "break the spatial profile": they live in areas with favorable spatial indicators yet face educational barriers comparable to their rural counterparts. The model has no pathway to identify them because the spatial features that capture indigenous disadvantage in rural settings do not activate in urban ones (Section 7). Sample caveat: $n = 89$ for urban other-indigenous students in the test set.

6.4 SHAP Interpretability

Table 7 and Figure 3 reveal how the model makes predictions, directly addressing RQ2. The top five SHAP features—age, nightlight z-score, working status, census literacy z-score, and poverty index z-score—are all spatial-structural variables. Identity features contribute minimally: the sex indicator (es_mujer) ranks 16th out of 25 features with a mean absolute SHAP value of only 0.003, and the nationality indicator (es_peruano) ranks 25th with effectively zero contribution, consistent with the $n = 27$ non-Peruvian sample producing no learnable signal.

Table 7. SHAP Feature Importance (Top 15)

Rank	Feature	Mean |SHAP|	LR Rank
1	Edad	0.1365	11
2	Intensidad de luces nocturnas (z)	0.0530	13
3	Trabaja	0.0483	6
4	Poblacion indigena del distrito (z)	0.0469	22
5	Tasa de alfabetismo del distrito (z)	0.0442	19
6	Indice de pobreza (z)	0.0340	9
7	Acceso a electricidad del distrito (z)	0.0331	15
8	Acceso a agua del distrito (z)	0.0323	18
9	Ingreso del hogar (log)	0.0318	14
10	Zona rural	0.0229	23
11	Tasa de desercion distrital (admin, z)	0.0160	24
12	Edad de secundaria (12+)	0.0147	4
13	Educacion de los padres (anos)	0.0092	25
14	Quintil de pobreza	0.0059	10
15	Otra lengua indigena	0.0056	1

SHAP computed on uncalibrated LightGBM; values in log-odds space

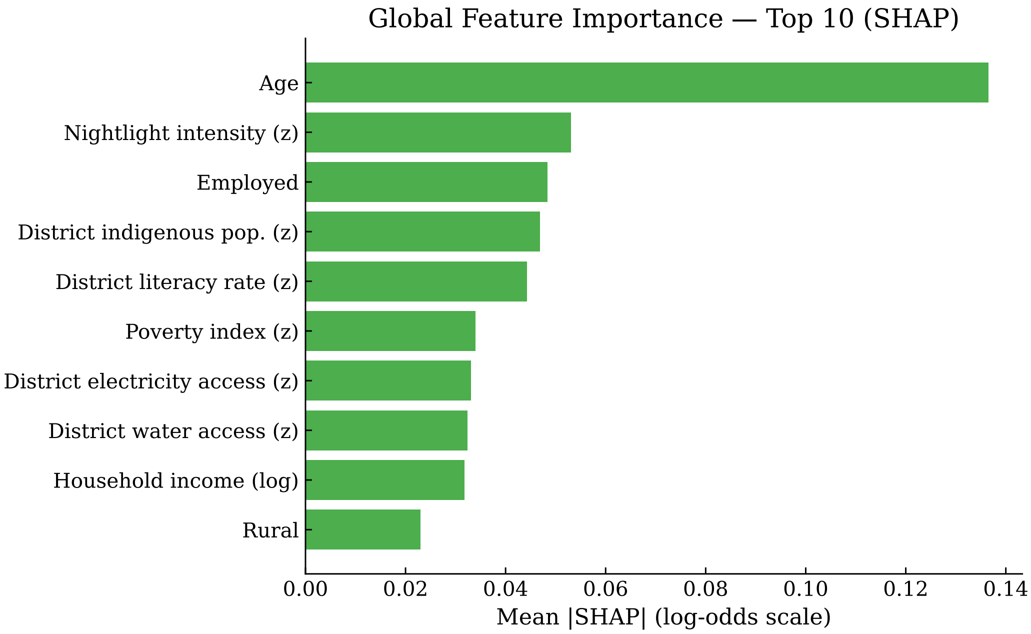

Fig. 3. Mean absolute SHAP values for the top 15 features. Age and spatial-structural features dominate, while identity features (language, sex) have minimal direct importance.

The top 5 SHAP features have zero overlap with the top 5 logistic regression features (dominated by indigenous language dummies): where logistic regression assigns large coefficients to identity features, LightGBM achieves similar discrimination through correlated spatial-structural features. The model encodes structural inequities without using identity features directly—removing protected attributes would not mitigate the documented disparities.

7 Discussion

7.1 The Spatial Proxy Mechanism

SHAP analysis reveals that nightlight intensity, district-level dropout rates, and census literacy rates collectively encode the spatial concentration of disadvantage, creating systematic blind spots for populations that do not match spatial stereotypes. Urban indigenous students exemplify this failure: they reside in areas with favorable spatial indicators yet face educational barriers comparable to their rural counterparts. This extends Perdomo et al.'s [29] finding that structural features predict dropout well by showing that reliance on such features creates predictable fairness failures at demographic intersections.

Feature ablation confirms this mechanism: removing district-level features reduces castellano FNR from 0.633 to 0.317, while the individual-only model slightly increases it to 0.649 (Table 11, Appendix B). Castellano speakers have the highest FNR in all three variants, confirming their invisibility is not an artifact of a particular feature set.

These findings suggest group-specific threshold adjustment could reduce castellano invisibility, and supplementary identification mechanisms could address the urban indigenous blind spot—though both approaches redistribute rather than eliminate errors, and their implementation requires consideration of operational costs and community consent [24].

We privilege FNR because a missed dropout represents irreversible harm. Equalizing FNR across language groups would require lowering the threshold for Spanish speakers, increasing their FPR. Whether this trade-off is acceptable depends on intervention cost: a teacher notification justifies elevated FPR; social worker home visits may not [9, 24]. The current model implicitly prioritizes low FPR for the majority at the cost of rendering their dropouts invisible.

8 Limitations

This paper audits a proxy model, not the actual Alerta Escuela system (Table 9, Appendix B).

Second, ENAHO's mother tongue variable (P300) captures language by self-report. Bilingual speakers may report Spanish as their mother tongue, potentially undercounting indigenous-language prevalence and understating the disparities we document. The true magnitude of language-based prediction disparities may be larger than our estimates.

Third, the 2020 wave (phone interviews, ~13,755 observations, 52% null attendance records) may not represent the same population as in-person survey years.

Fourth, intersectional subgroups have small samples: pooling val+test yields $n = 167$ urban other-indigenous students (FNR=0.69, 95% CI [0.39, 0.98]). A power analysis shows confirming FNR > 0.50 requires ~8 ENAHO years; detecting the gap versus castellano requires ~32 years. This methodological ceiling—survey data cannot produce significant intersectional results for subgroups with <6 dropouts per year—argues for opening SIAGIE administrative data.

Fifth, formal statistical guarantees for survey-weighted gradient boosting under complex designs remain an active research area [23].

9 Ethical Considerations

Positionality. The author is Peruvian, affiliated with UTEC and Genera (an edtech startup). I am not from the indigenous communities most affected by the disparities documented here; this work should be understood as an outsider's technical audit rather than a community-centered assessment.

Generative AI Disclosure. Portions of the data pipeline code and manuscript preparation were assisted by generative AI tools (Claude, Anthropic). The author was responsible for all research design, methodological decisions, result interpretation, and substantive writing. AI assistance was used for code implementation and editorial refinement.

Data Ethics. This study uses publicly available, de-identified survey data (ENAHO) released by INEI for research purposes. No individual students can be identified from the analysis, and no direct human subjects interaction was involved. The analysis operates exclusively on aggregate patterns in survey-weighted data.

10 Conclusion

This proxy equity audit of survey-derived dropout risk in Peru ($N = 150{,}135$; 2018–2023) identifies Spanish-speaking students—the demographic majority—as the group most systematically missed by the model (FNR = 0.633), with suggestive evidence that urban indigenous students face even higher miss rates at the intersection of language and geography (pooled FNR = 0.69, $n = 167$). These findings hold across five model families with different architectures, indicating that the disparity reflects data structure rather than algorithmic artifacts.

The proxy audit methodology demonstrates that independent algorithmic accountability is achievable using only publicly available survey data—an approach applicable wherever direct system access is unavailable. However, models that predict primarily through spatial-structural features create predictable blind spots at demographic intersections where geographic and social profiles diverge. Feature ablation confirms that spatial features drive the castellano invisibility pattern, while the full model's combination of spatial and individual features produces the worst FNR outcome for the majority group.

Our power analysis reveals a methodological ceiling: survey data fundamentally cannot produce statistically significant intersectional fairness results for subgroups contributing fewer than ~6 positive observations per year. Opening SIAGIE administrative data would enable both direct system evaluation and the statistical power that intersectional auditing demands. As educational early warning systems proliferate globally, the question of who decides the appropriate fairness trade-off—and who audits the systems making that trade-off—remains open.

A Supplementary Figures

B Supplementary Tables

References

[1] Melissa A. Adelman, Francisco Haimovich, Andrés Ham, and Emmanuel Vazquez. 2018. Predicting School Dropout with Administrative Data: New Evidence from Guatemala and Honduras. *Education Economics* 26, 4 (2018), 356–372. doi:10.1080/09645292.2018.1433127

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2623–2631.

[3] Ryan S. Baker and Aaron Hawn. 2022. Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education* 32, 4 (2022), 1052–1092. doi:10.1007/s40593-021-00285-9

[4] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (2016), 671–732.

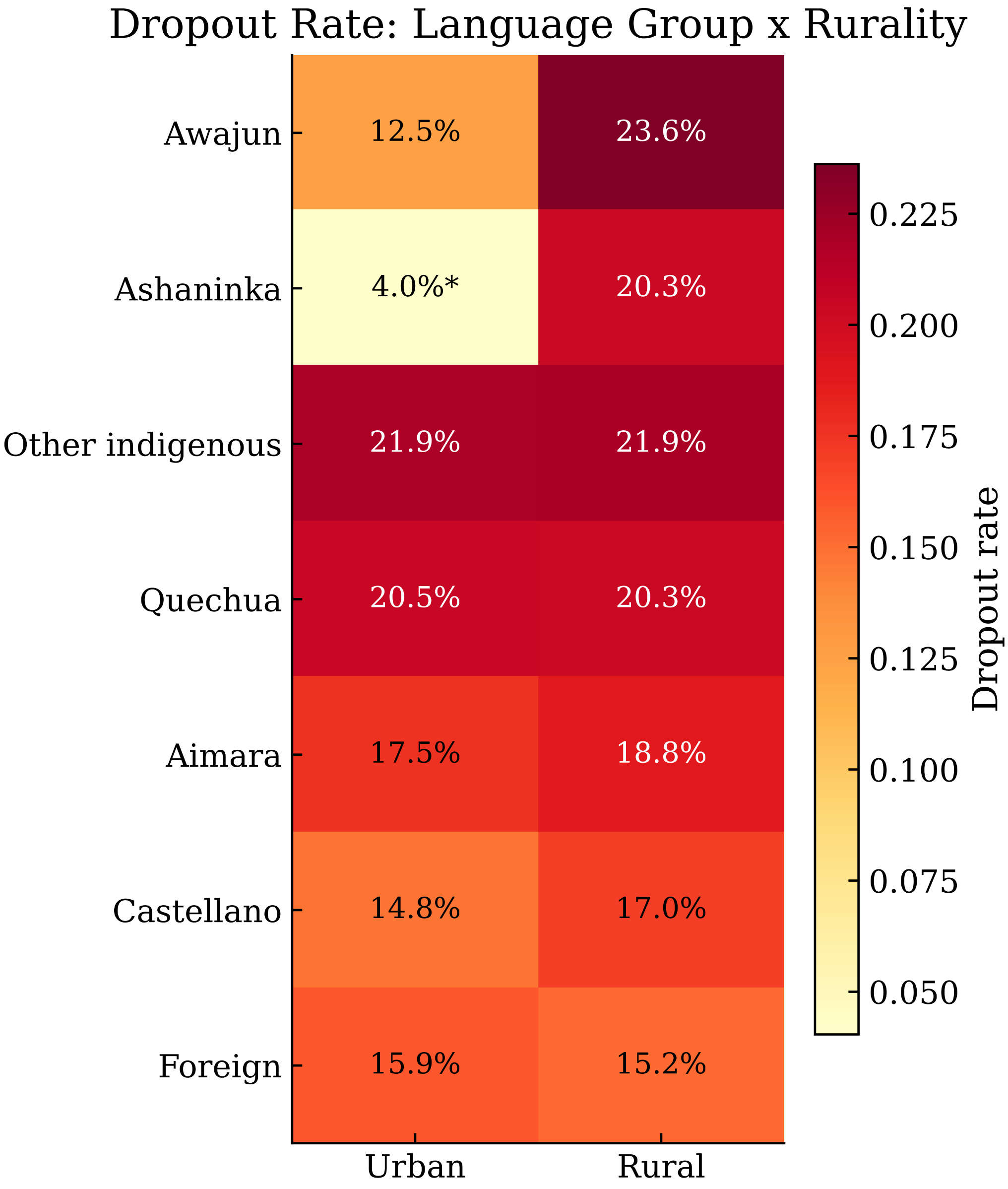

Fig. 4. Dropout rates by language group and rurality. Each cell shows the weighted dropout rate for the intersection of language and urban/rural geography.

[5] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. In *Microsoft Research Technical Report MSR-TR-2020-32*.

[6] Alex J. Bowers. 2010. Grades and Graduation: A Longitudinal Risk Perspective to Identify Student Dropouts. *The Journal of Educational Research* 103, 3 (2010), 191–207. doi:10.1080/00220670903382970

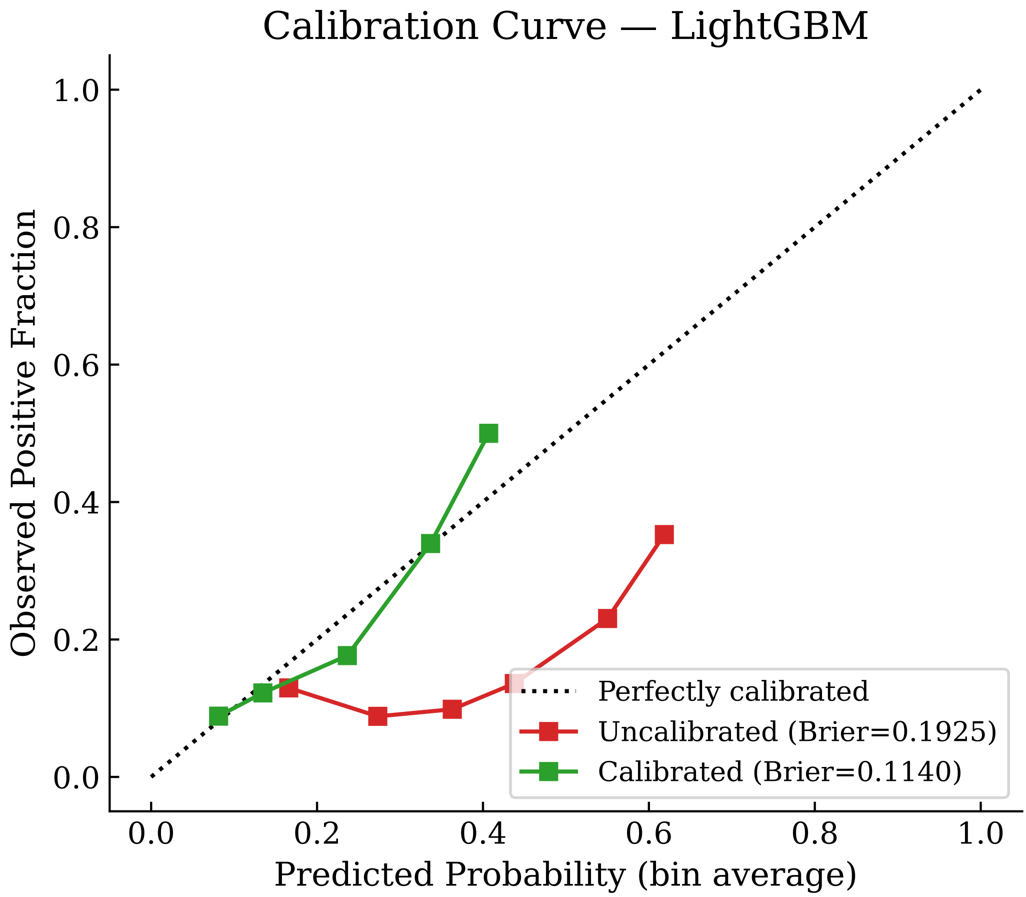

Fig. 5. Calibration plot comparing uncalibrated and Platt-calibrated LightGBM probabilities. Platt scaling reduces test Brier score by 40%.

[7] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of Machine Learning Research*, Vol. 81. PMLR, 77–91.

[8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.

[9] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163.

[10] Kimberlé Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum* 1989, 1 (1989), 139–167.

[11] Santiago Cueto, Gabriela Guerrero, Juan León, Ernesto Seguin, and Ismael Muñoz. 2009. Explaining and Overcoming Marginalization in Education: A Focus on Ethnic/Language Minorities in Peru. In *EFA Global Monitoring Report 2010 Background Paper*. UNESCO.

[12] Santiago Cueto, Alejandra Miranda, and Juan León. 2016. *Education Trajectories: From Early Childhood to Early Adulthood in Peru*. Country Report. Young Lives, University of Oxford.

[13] Josh Gardner, Christopher Brooks, and Ryan S. Baker. 2024. Debiasing Education Algorithms. *International Journal of Artificial Intelligence in Education* 34 (2024), 692–733. doi:10.1007/s40593-023-00389-4

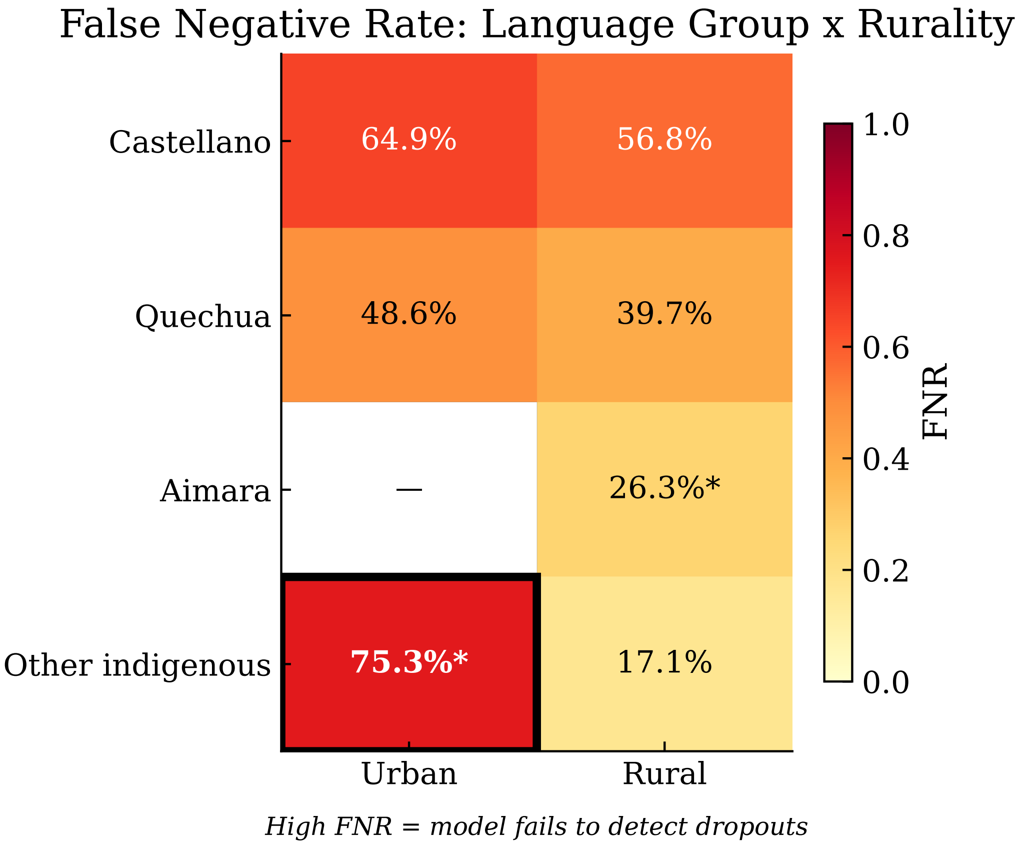

Fig. 6. FNR heatmap by language and rurality intersection. The darkest cell (other indigenous, urban) represents the group most missed by the model.

[14] Instituto Nacional de Estadística e Informática. 2023. *Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza (ENAHO): Metodología y Documentación Técnica.* Technical Report. INEI, Lima, Perú. https://www.inei.gob.pe/.

[15] Marzieh Karimi-Haghighi, Carlos Castillo, Albert Diaz-Guilera, and Sergio Luján-Mora. 2021. Predicting Early Dropout: Calibration and Algorithmic Fairness Considerations. *arXiv preprint arXiv:2103.09068* (2021).

[16] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[17] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Vol. 80. PMLR, 2564–2572.

[18] René F. Kizilcec and Hansol Lee. 2022. Algorithmic Fairness in Education. In *The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates*, Wayne Holmes and Kaśka Porayska-Pomsta (Eds.). Routledge, 174–202. doi:10.4324/9780429329067-10

[19] Jared E. Knowles. 2015. Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin. *Journal of Educational Data Mining* 7, 3 (2015), 18–67. doi:10.5281/zenodo.3554726

[20] Himabindu Lakkaraju, Everaldo Aguiar, Carl Shan, David Miller, Nasir Bhanpuri, Rayid Ghani, and Kecia L. Addison. 2015. 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*. ACM, 1909–1918. doi:10.1145/2783258.2788620

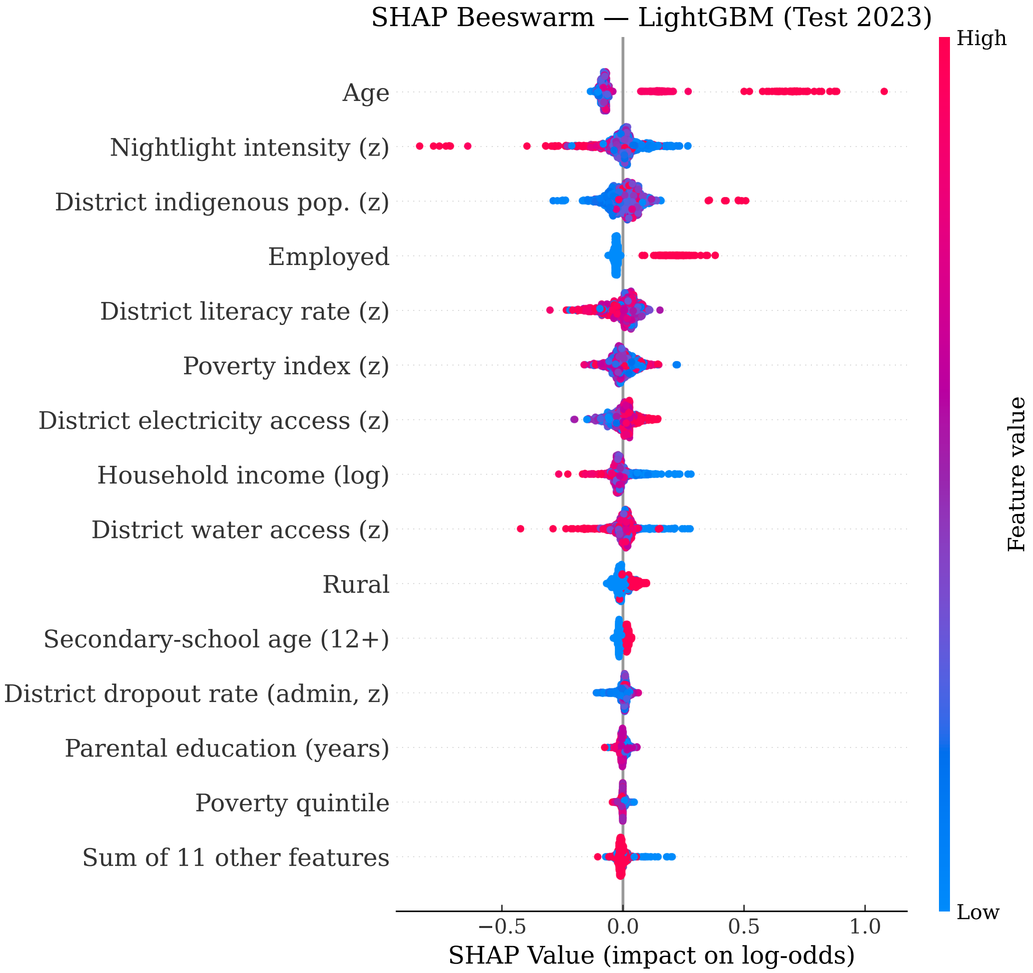

Fig. 7. SHAP beeswarm plot showing feature value distributions and their impact on predictions. Red indicates high feature values; blue indicates low values.

[21] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence* 2, 1 (2020), 56–67.

[22] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[23] Nathaniel MacNell, Lydia Feinstein, Jesse Wilkerson, Päivi M. Salo, Samantha A. Molsberry, Michael B. Fessler, Peter S. Thorne, Alison A. Motsinger-Reif, and Darryl C. Zeldin. 2023. Implementing Machine Learning Methods with Complex Survey Data: Lessons Learned on the Impacts of Accounting Sampling Weights in Gradient Boosting. *PLOS ONE* 18, 1 (2023), e0280387. doi:10.1371/journal.pone.0280387

[24] Brian McMahon, Nathan R. Todd, Amy Martinez, Chelsey Coker, Chia-Fang Sheu, Jason Washburn, and Sachin Shah. 2020. Re-envisioning the Purpose of Early Warning Systems: Shifting the Mindset from Student Identification to Meaningful Prediction and Intervention. *Review of Education* 8, 1 (2020), 266–301. doi:10.1002/rev3.3183

[25] Ministerio de Educación del Perú. 2022. Estadística de la Calidad Educativa (ESCALE). https://escale.minedu.gob.pe/. Accessed: 2026-02-01.

Table 8. Logistic Regression Coefficients (All 25 Features)

Feature	Coefficient	Odds Ratio	Dir.
Otra lengua indigena	0.7880	2.199	↑
Lengua extranjera	0.5760	1.779	↑
Lengua quechua	0.4713	1.602	↑
Edad de secundaria (12+)	-0.4378	0.645	↓
Lengua aimara	0.3465	1.414	↑
Trabaja	0.3460	1.413	↑
Nacionalidad peruana	0.2985	1.348	↑
Lengua castellana	0.2952	1.343	↑
Indice de pobreza (z)	0.2525	1.287	↑
Quintil de pobreza	-0.2100	0.811	↓
Edad	0.1201	1.128	↑
Tiene discapacidad	0.1148	1.122	↑
Intensidad de luces nocturnas (z)	-0.1074	0.898	↓
Ingreso del hogar (log)	-0.1073	0.898	↓
Acceso a electricidad del distrito (z)	0.0917	1.096	↑
Region Selva	-0.0772	0.926	↓
Region Sierra	-0.0722	0.930	↓
Acceso a agua del distrito (z)	-0.0536	0.948	↓
Tasa de alfabetismo del distrito (z)	-0.0516	0.950	↓
Sexo femenino	-0.0438	0.957	↓
Beneficiario JUNTOS	-0.0335	0.967	↓
Poblacion indigena del distrito (z)	0.0216	1.022	↑
Zona rural	-0.0100	0.990	↓
Tasa de desercion distrital (admin, z)	-0.0031	0.997	↓
Educacion de los padres (anos)	-0.0018	0.998	↓

[26] Ministerio de Educación del Perú. 2023. Alerta Escuela: Sistema de Alerta Temprana para la Prevención de la Deserción Escolar. https://www.gob.pe/minedu. Accessed: 2026-02-01.

[27] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.

[28] Chenguang Pan and Zhou Zhang. 2024. Examining the Algorithmic Fairness in Predicting High School Dropouts. In *Proceedings of the 17th International Conference on Educational Data Mining (EDM 2024)*. Atlanta, GA, 207–214.

[29] Juan C. Perdomo, Tolani Britton, Moritz Basu, Jon Kleinberg, and Sendhil Mullainathan. 2025. Difficult Lessons on Social Prediction from Wisconsin Public Schools. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

[30] John C. Platt. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers* (1999), 61–74.

[31] William Villegas-Ch, Aracely Arias-Navarrete, and Xavier Palacios-Pacheco. 2023. Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic, Socioeconomic, and Equity Factors through Machine Learning and Survival Analysis Methods in the Latin American Context. *Education Sciences* 13, 2 (2023), 154. doi:10.3390/educsci13020154

Table 9. ENAHO vs. SIAGIE Feature Availability. SIAGIE columns are inferred from public documentation [25, 26]; we have not accessed SIAGIE records directly.

Feature Category	ENAHO (this study)	SIAGIE (Alerta Escuela, inferred)
Demographics	Age, sex, mother tongue, nationality (self-report)	Name, DOB, sex, grade, school enrollment (administrative)
Economic	Poverty index, household expenditure, poverty quintile	Free lunch eligibility (inferred); no income/expenditure
Geographic	Department, district, natural region	School location, district code
Attendance/School	Current enrollment (annual self-report)	Daily attendance records, grade history
Longitudinal	6 annual waves pooled (cross-section per year)	Continuous multi-year student trajectory
Data characteristics		
Coverage	150,135 school-age obs (ages 6–17)	∼2M enrolled students/year (estimated)
Unit of observation	Household survey respondent	Administrative student record
Frequency	Annual survey wave	Continuous / daily
Publicly accessible	Yes (INEI, open data)	No (MINEDU internal use only)

Table 10. False Negative Rate by Language Group Across Five Model Families. Aimara group ($n = 76$) shows instability (MLP FNR=0.830); cross-architecture consistency claim is scoped to the castellano vs. indigenous pattern.

Language Group	LR	LightGBM	XGBoost	RF	MLP
Castellano	0.584	0.633	0.613	0.549	0.666
Quechua	0.192	0.416	0.284	0.259	0.525
Otros indígenas	0.065	0.216	0.159	0.216	0.397
Aimara*	0.288	0.263	0.288	0.192	0.830

* $n = 76$; Aimara MLP FNR=0.830 is a small-sample outlier.

Cross-architecture consistency claim applies to castellano vs. indigenous pattern only.

Table 11. FNR by Language Group Under Feature Ablation. "Individual only" removes all 7 district-level spatial features; "Spatial only" removes all 18 individual/household features. Each variant uses its own optimal threshold (max weighted F1 on validation).

Language Group	Full Model (25 features)	Individual Only (18 features)	Spatial Only (7 features)
Castellano	0.633	0.649	0.317
Quechua	0.416	0.192	0.160
Aimara	0.263	0.288	0.188
Other indigenous	0.216	0.136	0.131
Val PR-AUC	0.262	0.250	0.176