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

This paper does not audit Peru's Alerta Escuela early warning system directly—we have not accessed its predictions, training data, or operational feature set. Instead, we construct a proxy dropout prediction model from publicly available ENAHO survey data targeting the same school-age population, and use it to examine what fairness disparities can emerge from survey-derived dropout risk modeling. Using six years of nationally representative data (2018–2023, $N = 150{,}135$), we train five model families (logistic regression, LightGBM, XGBoost, random forest, and MLP) and conduct a comprehensive equity audit across language, geography, poverty, and sex dimensions. The calibrated LightGBM model achieves a test PR-AUC of 0.236 with a top-decile lift of 2.54× over the baseline prevalence, demonstrating meaningful predictive signal. The model exhibits a false negative rate (FNR) of 63% for Spanish-speaking students but only 22% for indigenous-language speakers—revealing a surveillance–invisibility axis where indigenous students are over-flagged while the majority of Spanish-speaking dropouts are missed. This FNR rank order is consistent across all five model families. SHAP analysis shows the model predicts through spatial-structural features (age, nightlights, literacy rates, poverty) rather than identity features directly. The starkest disparity emerges at the intersection of language and urbanicity: urban indigenous students face a 75% FNR ($n = 89$; 95% CI [0.211, 1.000]). Our contributions are twofold: (1) a proxy audit framework demonstrating that independent algorithmic accountability is achievable using only public data, and (2) empirical documentation of equity gaps in Peruvian dropout risk prediction that any system operating on similar demographic structure may exhibit.

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 [27]. Such systems promise efficiency and early intervention, but their algorithmic fairness properties remain almost entirely unaudited. A growing body of research has demonstrated that predictive models can systematically disadvantage marginalized groups—encoding structural inequities into automated decisions that affect millions of students [5, 28].

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 [4, 14, 18]. This gap is particularly consequential in countries like Peru, where the axes of disadvantage—mother tongue, geography, poverty—differ fundamentally from those studied in the Global North.

Peru is a multilingual country where approximately 16% of the population speaks an indigenous language as their mother tongue. Indigenous-language speakers face persistent educational inequities rooted in colonial legacies, geographic isolation, and inadequate bilingual education—only 37% of indigenous students attend schools with bilingual instruction [12, 13]. Peru's Encuesta Nacional de Hogares (ENAHO), conducted annually by the Instituto Nacional de Estadística e Informática (INEI), provides nationally representative data that enables analysis of these disparities [15]. Because the actual SIAGIE administrative data used by Alerta Escuela is not publicly accessible,

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.

we construct a proxy replication of an Alerta Escuela–style dropout prediction model using ENAHO survey data spanning 150,135 student-year observations across six years (2018–2023).

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 [17, 32], evaluate fairness across seven demographic dimensions and three intersections using the fairlearn framework [6], and apply SHAP TreeExplainer to decompose predictions into feature-level contributions [23]. 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: the model over-flags indigenous-language students (low false negative rate but high false positive rate) while missing the majority of Spanish-speaking dropouts (high FNR, low FPR). The starkest disparity emerges at the intersection of language and urbanicity, where the model fails to identify most dropouts in a specific subgroup. SHAP analysis shows the model predicts through spatial-structural proxy features rather than identity features, and the pattern holds across all five model families.

Our contributions are:

- A proxy equity audit framework demonstrating that independent algorithmic accountability is achievable using only publicly available survey data, without access to the system's training data or operational predictions—enabling accountability where direct system access is unavailable.
- A comprehensive fairness audit spanning seven demographic dimensions and three intersections with survey-weighted metrics, demonstrating that intersectional analysis reveals disparities hidden by single-axis evaluation—urban indigenous students emerge as the most systematically missed group only when language and urbanicity are crossed [8, 11].
- Evidence that dropout prediction models encode structural inequities through spatial-structural proxy features rather than through explicit use of protected attributes, with implications for fairness interventions.
- An open-source, replicable audit framework—code, data pipeline, and analysis are publicly available to enable similar audits of educational EWS in other contexts.

What this paper does not claim is equally important. We have not accessed Alerta Escuela's actual predictions, training data, or operational feature set. This paper audits a proxy model built from ENAHO survey data targeting the same school-age population. We do not claim that the actual Alerta Escuela system exhibits the specific disparities documented here, that SIAGIE-trained models would produce identical fairness profiles, or that the Ministry of Education's system is biased in this way. Our findings demonstrate what disparities *can* emerge from survey-derived dropout prediction in Peru's demographic context—a class of model to which Alerta Escuela-style systems belong, but about which we make no specific operational claims.

2 Related Work

2.1 Early Warning Systems in Education

Dropout early warning systems have evolved substantially over the past two decades. Bowers [7] established foundational indicators—grades, GPA, and course failures—as predictors of dropout risk in a longitudinal study of US high school students. As machine learning methods matured, researchers developed increasingly sophisticated systems: Lakkaraju et al. [20] introduced an ML framework for identifying at-risk K-12 students, demonstrating that ensemble methods could substantially outperform traditional indicator thresholds. Knowles [19] deployed the first statewide ML-based dropout EWS in Wisconsin, covering over 225,000 students with administrative data—a scale comparable to Peru's Alerta Escuela.

Subsequent work has expanded both the methods and the contexts. Baker et al. [3] applied logistic regression to attendance, grades, and disciplinary data in diverse US school districts, while Lee and Chung [21] established the pattern of temporal train-test splits that mirrors real-world deployment—a design we adopt. In the developing-country context most relevant to our work, Adelman et al. [1] used administrative data to predict dropout in Guatemala and Honduras, correctly identifying 80% of eventual dropouts.

Two recent contributions provide critical framing for our study. McMahon et al. [25] argue that EWS should shift from pure identification to meaningful prediction-plus-intervention, questioning whether flagging students as "at-risk" without adequate support mechanisms constitutes a net benefit. Most provocatively, Perdomo et al. [31] evaluate Wisconsin's deployed EWS over a decade and argue that structural features predict dropout as well as individual risk scores—a finding our SHAP analysis directly corroborates. Our paper extends this critical tradition by auditing an EWS-style model in a context where deployment occurs but fairness evaluation does not.

2.2 Algorithmic Fairness in Education

Kizilcec and Lee [18] identified that fairness audits, standard in criminal justice and hiring, remained rare in deployed educational systems. Baker and Hawn [4] catalogued known biases across educational applications and introduced "slice analysis" for disaggregated evaluation. Chouldechova [10] proved that no classifier can simultaneously satisfy calibration, equal FNR, and equal FPR across groups with different base rates—an impossibility result our FNR-FPR trade-off directly illustrates.

Pan and Zhang [30] examined fairness in US high school dropout prediction but without survey weights or intersectional analysis. Karimi-Haghighi et al. [16] combined calibration and fairness in dropout prediction but without survey-weighted methodology. Gardner, Brooks, and Baker [14] found most debiasing studies focus on gender and race in US/European contexts. Our paper fills this gap: a comprehensive proxy audit in a developing-country, multilingual context using survey-weighted analysis across seven dimensions and three intersections.

2.3 Fairness in Latin American Educational AI

Latin American education systems face structural inequalities rooted in colonial histories, geographic barriers, and linguistic diversity [34]. In Peru specifically, indigenous-language speakers experience persistent educational disadvantage. Cueto et al. [12] documented how ethnic and language minorities in Peru are systematically marginalized in education, finding that indigenous-language students receive lower-quality instruction and face cultural barriers to school engagement. Cueto, Miranda, and León [13] traced education trajectories from early childhood through adolescence using Young Lives longitudinal data, reporting that only 37% of indigenous students attend bilingual schools despite legal mandates for intercultural bilingual education.

Machine learning is increasingly applied to dropout prediction in the region. Adelman et al. [1] demonstrated effective dropout prediction in Guatemala and Honduras using administrative data, and Villegas-Ch et al. [35] evaluated ML approaches for higher education dropout in a Latin American context, finding that socioeconomic variables dominate prediction. Notably, none of these studies conducted fairness audits of their prediction systems. Our paper provides the first such audit for a Peruvian educational prediction system, examining whether the demographic disparities documented by Cueto and colleagues are reproduced—or amplified—by algorithmic prediction.

2.4 Intersectionality in ML Fairness

Crenshaw [11] established that single-axis analysis systematically misses compound marginalization. Buolamwini and Gebru [8] demonstrated this computationally: facial recognition error rates of 34.7% for darker-skinned females vs. 0.8% for lighter-skinned males were invisible in single-axis analysis. Our intersectional analysis operationalizes these frameworks in an educational context: the language × urbanicity intersection that reveals urban indigenous students' FNR of 0.753 parallels the Gender Shades finding—a group invisible to single-axis evaluation that faces the most extreme prediction errors.

3 Data

Peru has approximately 8 million school-age children, and dropout remains a persistent challenge—particularly in rural areas and among indigenous-language communities. The Ministry of Education (MINEDU) operates Alerta Escuela, which uses data from the Sistema de Información de Apoyo a la Gestión de la Institución Educativa (SIAGIE) to flag students at risk of dropout [26, 27]. Because SIAGIE administrative records are not publicly accessible, we use ENAHO survey data as a proxy to construct and audit an Alerta Escuela–style prediction model.

The comparison in Table 1 highlights a key limitation of the proxy approach: SIAGIE contains daily attendance records, multi-year student trajectory, and grade history that ENAHO does not capture. Our proxy model predicts from annual cross-sectional survey data, missing the longitudinal signal that likely improves the actual system's predictive accuracy. However, the survey dimensions available in ENAHO—mother tongue, poverty, geography—are precisely those needed to study equity disparities, and these dimensions are either absent from or not publicly reported for SIAGIE-based models.

We use Peru's Encuesta Nacional de Hogares (ENAHO), a nationally representative household survey conducted annually by the Instituto Nacional de Estadística e Informática (INEI) [15]. ENAHO employs a multi-stage, stratified sampling design covering all 25 departments of Peru, with survey weights (FACTOR07) that account for the complex sampling structure and enable nationally representative inference. We extract data from Module 200 (demographic characteristics) and Module 300 (education), joining them by household and person identifiers.

Our analysis pools six annual waves (2018–2023) covering school-age children aged 6–17, yielding 150,135 individual-year observations after data cleaning. The 2020 wave is notably affected by the COVID-19 pandemic: INEI conducted phone interviews rather than in-person visits, resulting in a reduced sample of approximately 13,755 observations (compared to approximately 25,000 in a typical year) and 52% null values in the education attendance variable (P303), which were dropped. 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.

Table 2 summarizes the sample across key demographic dimensions. 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.

Table 1. ENAHO vs. SIAGIE Feature Availability. SIAGIE columns are inferred from public documentation [26, 27]; we have not accessed SIAGIE records directly. This comparison documents the features *not* available in our proxy model that may be present in the actual system.

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)

The sample spans all three major geographic regions: Costa (coast), Sierra (highlands), and Selva (Amazon lowlands).

Table 3 reveals substantial disparities in weighted dropout rates across language groups. Indigenous-language speakers face rates 34% higher than Spanish speakers on average. The Otros indígenas group exhibits the highest dropout rate at 0.219, followed by Awajun at 0.205, compared to 0.153 for Castellano speakers—a gap that persists even after accounting for geographic and socioeconomic differences. For the fairness analysis (Section 6), Ashaninka and Awajun are grouped under “Otros indígenas” due to small per-group sample sizes that would yield unreliable metric estimates; Extranjero speakers are excluded from language-dimension analysis given the proxy model’s focus on indigenous–Spanish disparities. This consolidation accounts for the difference between the stated test set size (n = 25,635) and the language fairness table sum (n = 25,592): the 43 Extranjero students in the 2023 test set are excluded from Table 8.

Table 4 and Figure 1 show how dropout rates vary by region, poverty level, and their interactions with language. The Sierra and Selva regions exhibit higher dropout rates than the Costa, and a largely monotonic poverty gradient is visible (with a minor reversal between Q2 and Q3): the poorest quintile has substantially higher dropout rates than the wealthiest. Figure 1 further reveals that the interaction of language and rurality produces disparities that exceed what either dimension alone would suggest.

Table 2. 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

Table 3. 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 4. 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]

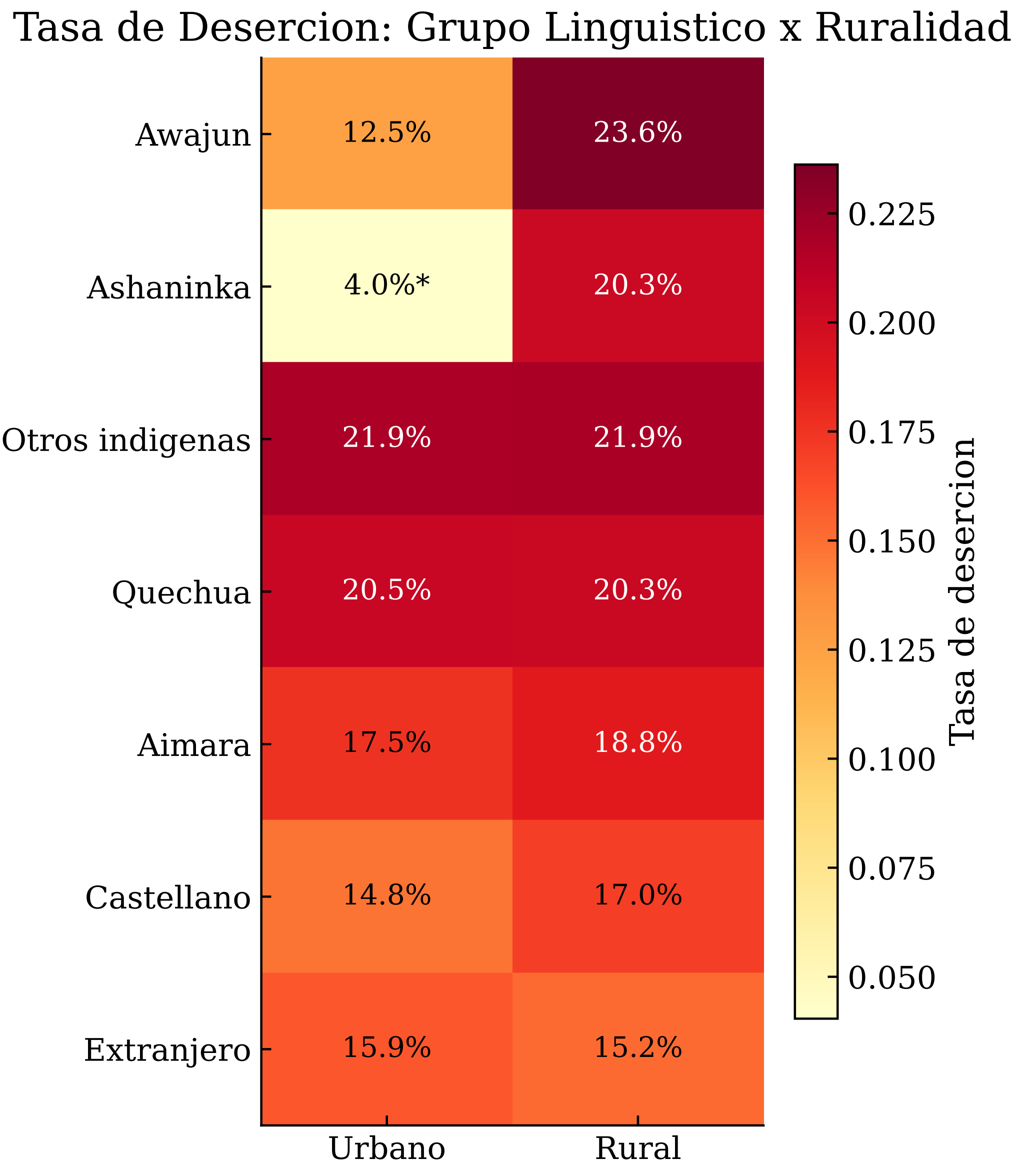

Fig. 1. Dropout rates by language group and rurality. Each cell shows the weighted dropout rate for the intersection of language and urban/rural geography.

In addition to individual and household variables from ENAHO, we merge district-level spatial features to capture contextual effects. Census data provides district population and literacy rate z-scores. Nightlight intensity, measured by satellite remote sensing, serves as a proxy for local economic activity and infrastructure. Administrative records from MINEDU provide district-level primaria (primary) and secundaria (secondary) completion rates. Merge rates are high: 100% for administrative and census data, and 95.9% for nightlight data, with 44 districts (1.53%) having primaria but no secundaria administrative records.

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 6 lists all 25 features with their logistic regression coefficients. Nightlight z-score nulls (4.1%) are imputed with 0.0 [24]; poverty quintiles are constructed using FACTOR07-weighted quantiles.

4.2 Model Selection and Training

We compare five model families chosen for complementary purposes. *Logistic regression* provides interpretable coefficients and odds ratios; we fit both a scikit-learn implementation for prediction and a statsmodels GLM with Binomial family for statistical inference. *LightGBM* [17] serves as the primary predictive model, leveraging gradient boosting's ability to capture nonlinearities and feature interactions. *XGBoost* [9] provides a second gradient boosting implementation for algorithm-independence checking. *Random Forest* extends the ensemble comparison beyond boosting to bagging. *MLP* (multilayer perceptron) provides a neural network baseline with fundamentally different inductive biases from tree-based models. If fairness findings hold across all five families, 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

Gradient boosted trees produce probability estimates that are often poorly calibrated, particularly when class weighting is applied to handle imbalanced outcomes [29]. We apply Platt scaling [32] to the LightGBM model's raw probability outputs, fitting a sigmoid function on the validation set. This reduces the validation Brier score from 0.186 to 0.116—a 38% improvement—confirming that calibration is critical for models with scale_pos_weight adjustments. The Platt scaling parameters ($A = -6.236$, $B = 4.443$) compress the raw probability range, with calibrated probabilities reaching a maximum of approximately 0.43.

4.4 Fairness Evaluation Framework

Fairness evaluation uses the fairlearn framework [6] to compute disaggregated metrics across seven demographic dimensions (language, natural region, rurality, poverty quintile, sex, nationality, age group) and three intersections (language × rurality, language × poverty quintile, language × region). For each subgroup, we compute four metrics: false negative rate (FNR, the proportion of actual dropouts missed by the model), false positive rate (FPR, the proportion of non-dropouts incorrectly flagged), precision, and PR-AUC. All metrics are computed with survey weights. This framework operationalizes the "slice analysis" approach advocated by Baker and Hawn [4] and aligns with the audit methodology of Saleiro et al. [33].

The choice of FNR as a primary fairness metric reflects its direct operational interpretation: a high FNR means the system fails to identify students who will drop out. From an equity perspective, FNR disparities indicate which populations are systematically rendered invisible to the early warning system. We complement FNR with FPR to capture the surveillance–invisibility trade-off that Chouldechova's [10] impossibility theorem predicts will arise when base rates differ across groups.

SHAP TreeExplainer [22] provides feature-level interpretability, decomposing each prediction into additive feature contributions. We compute SHAP values on the raw (uncalibrated) LightGBM model, as TreeExplainer requires direct access to the tree structure. SHAP interaction values are computed on a 1,000-row subsample of the test set.

5 Results

Table 5. 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

Table 5 compares five model families. LightGBM, XGBoost, and RF achieve near-identical validation PR-AUC (0.262, 0.263, and 0.261 respectively). MLP achieves PR-AUC of 0.238, lower than the tree-based ensembles as is typical on structured tabular data [17]. The calibrated LightGBM model achieves a test PR-AUC of 0.236, with a validation-test gap of 0.023 from unrounded values (well within the 0.07 threshold that would indicate concerning generalization failure). The calibrated Brier score of 0.112 on the 2023 test set (a 40% reduction from uncalibrated) confirms that calibration is essential for interpreting predicted probabilities as actual dropout risks.

Table 6 shows the logistic regression coefficients. Indigenous language variables dominate the linear model ("other indigenous" odds ratio = 2.20), contrasting sharply with the SHAP analysis of tree-based models in Section 6, where spatial-structural features dominate. This paradigm difference (zero overlap in top-5 features between linear and tree-based models) demonstrates that feature "importance" depends on model family.

Figure 2 visually confirms the algorithm independence: LightGBM and XGBoost PR curves largely overlap. Figure 3 shows that Platt scaling corrects the systematic overestimation caused by scale_pos_weight, producing probabilities that match observed dropout rates.

5.1 Predictive Validity

Before examining fairness properties, we establish that the model has meaningful predictive signal. 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%

Table 6. 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	↓

of students contains 34.2% actual dropouts—a lift of 2.54× over the 13.4% baseline (Figure 4). This decile-level concentration of risk confirms that the model's predictions are meaningful, not random.

Figure 4 shows calibration by prediction decile. The model is well-calibrated across the score range (mean absolute calibration error = 0.018), meaning a predicted probability of 0.30 corresponds to approximately 30% actual dropout in that decile. The Brier Skill Score of 0.040 for the calibrated model is positive, confirming it outperforms the prevalence baseline. Among the uncalibrated models, LR, XGBoost, and RF have negative Brier Skill Scores due to scale_pos_weight distorting raw probabilities. MLP achieves a marginal positive BSS of 0.012 because its sigmoid output is not affected by scale_pos_weight; nevertheless, Platt scaling on LightGBM remains the only well-calibrated model we recommend for probability-based decisions.

We acknowledge that PR-AUC of 0.236 is a modest absolute value. However, low absolute PR-AUC does not invalidate differential FNR findings: a model can be modestly predictive overall while exhibiting systematic and substantial differences in prediction errors across demographic subgroups. The fairness analysis that follows documents those differences across five model families with different architectures—robustness across architectures provides stronger evidence than any single model's absolute performance.

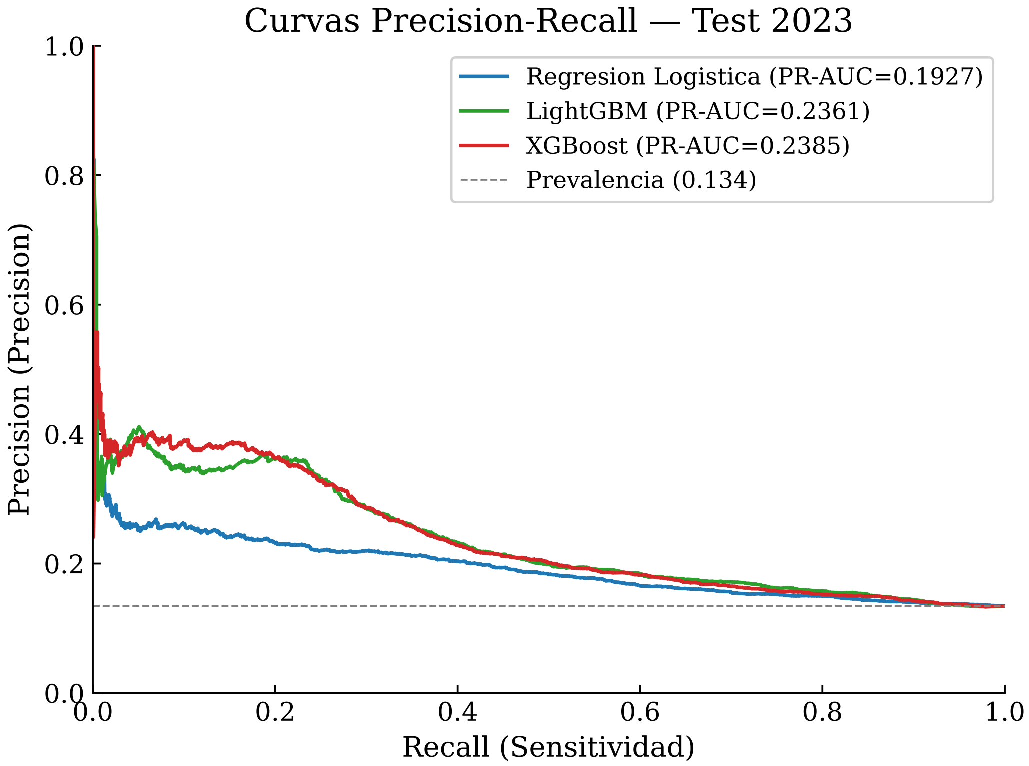

Fig. 2. Precision-Recall curves for three of the five model families on the 2022 validation set. LightGBM and XGBoost curves largely overlap; RF and MLP curves are omitted for visual clarity (see Table 5 for all five).

5.2 Algorithm Independence

Table 7 extends the algorithm-independence check to five model families (LR, LightGBM, XGBoost, RF, MLP) using the FNR disparity that is central to our fairness findings. Across all five architectures, Castellano speakers consistently show higher FNR than Quechua and other-indigenous speakers—the rank order that defines our surveillance–invisibility finding is not an artifact of the LightGBM implementation.

Table 7. False Negative Rate by Language Group Across Five Model Families. Aimara group ($n = 76$) shows instability (MLP FNR=0.830); algorithm independence 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.

Algorithm independence claim applies to castellano vs. indigenous pattern only.

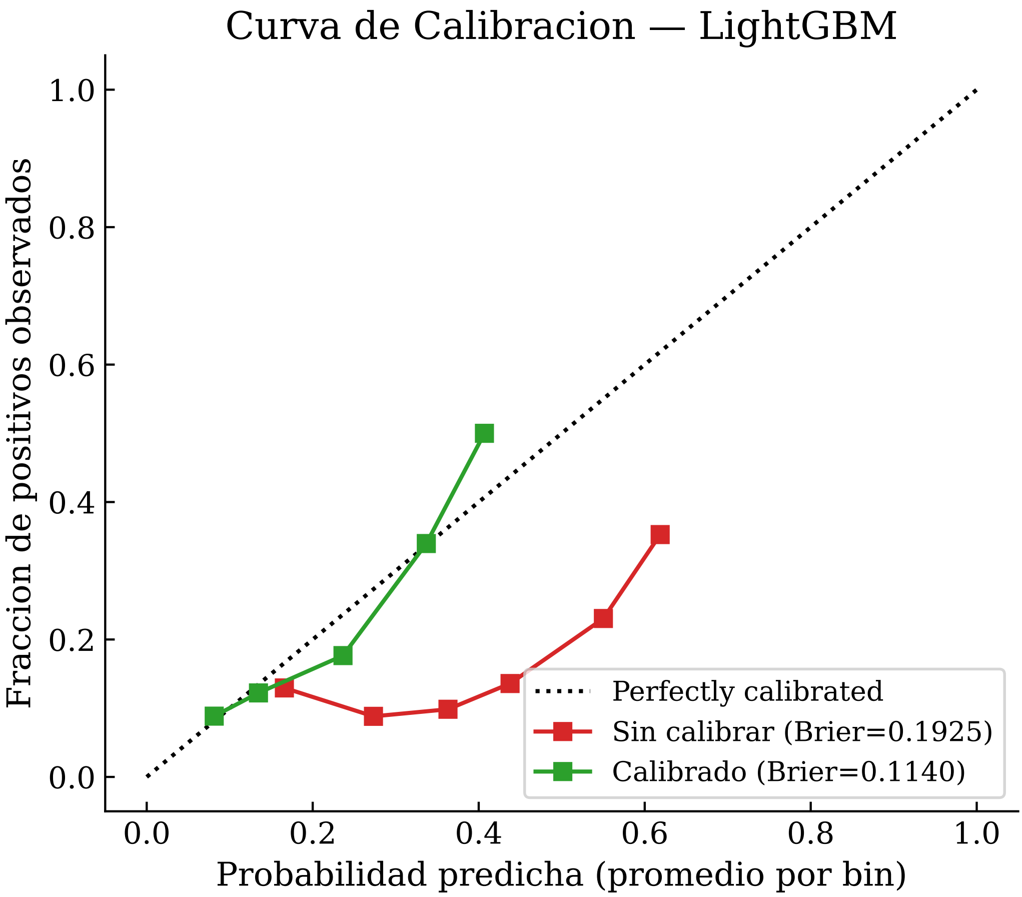

Fig. 3. Calibration plot comparing uncalibrated and Platt-calibrated LightGBM probabilities. Platt scaling reduces test Brier score by 40%.

The absolute FNR values vary across architectures—LR shows a narrower range (0.065–0.584) than LightGBM (0.216–0.633)—but the ordinal pattern is consistent: castellano FNR exceeds quechua FNR, which exceeds other-indigenous FNR in all five models. This consistency across architectures with different inductive biases (linear vs. gradient boosting vs. neural network) indicates that the disparity reflects data structure, not modeling artifacts. We note that the MLP Aimara FNR of 0.830 is an outlier driven by the small sample (n = 76) and should not be interpreted as a substantive finding.

6 Fairness Analysis

6.1 Language Dimension: The Surveillance–Invisibility Axis

Table 8 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,

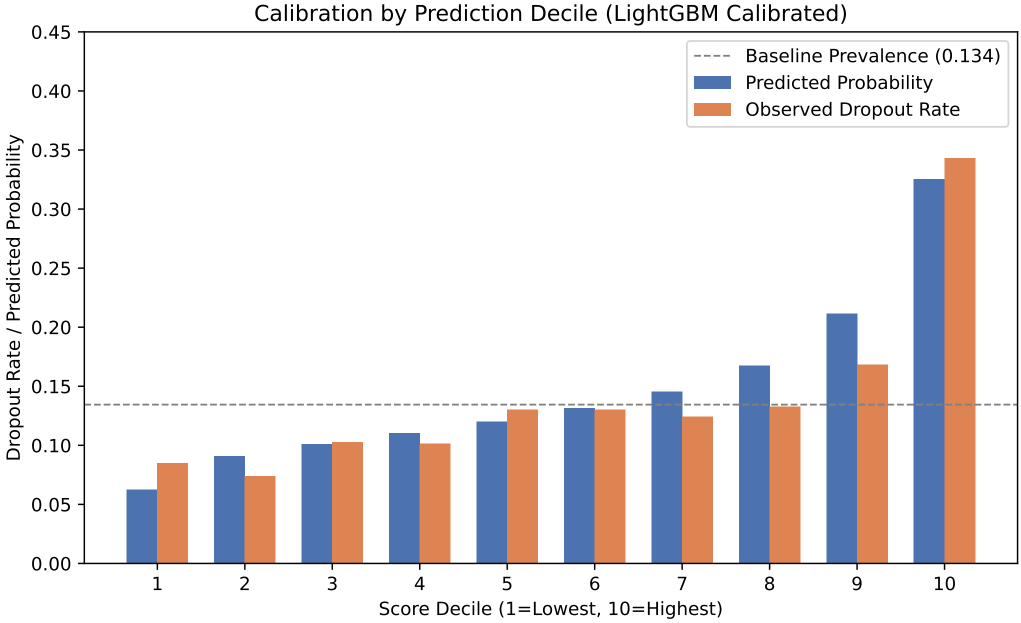

Fig. 4. Calibration by prediction decile for the LightGBM calibrated model. Bars show predicted probability (blue) and observed dropout rate (orange) per decile. Mean absolute calibration error = 0.018, indicating well-calibrated predictions. Baseline dropout prevalence = 0.134 (dashed line).

Table 8. 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).

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 FNR-FPR relationship is not a model bug but the mathematical consequence of Chouldechova’s [10] impossibility result applied to groups with different base rates. Indigenous-language speakers have higher baseline dropout rates, so a model trained to minimize overall prediction error will flag them more aggressively. The result is a systematic redistribution of prediction errors: indigenous communities bear the burden of false alarms (surveillance) while Spanish-speaking dropouts bear the burden of being missed (invisibility).

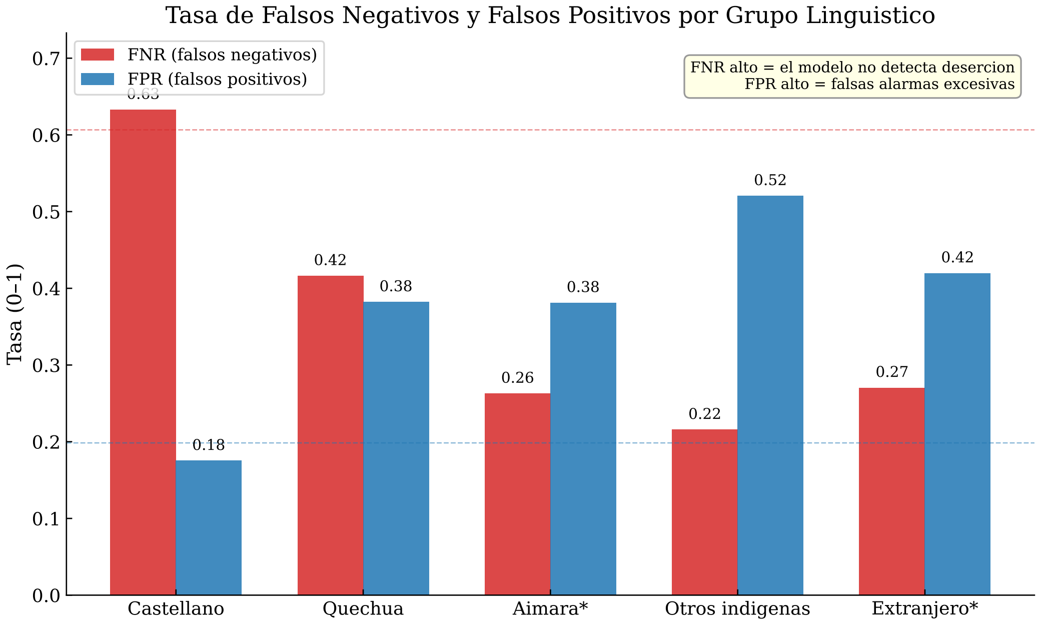

Fig. 5. FNR and FPR by language group. The inverse relationship between FNR and FPR reveals the surveillance–invisibility trade-off.

Figure 5 visualizes this trade-off. The inverse relationship between FNR and FPR across language groups forms a clear axis: as FNR decreases (better detection), FPR increases (more false alarms), with indigenous-language groups clustered at the high-detection/high-surveillance end and Spanish speakers at the low-detection/low-surveillance end.

6.2 Other Demographic Dimensions

The fairness analysis extends beyond language to six additional dimensions. *Region:* The Selva (Amazon) and Sierra (highlands) regions show lower FNR than the Costa (coast)—the model detects rural and remote dropouts more effectively, likely because these students match the spatial profile most strongly associated with dropout risk. However, a calibration gap exists: students predicted as "high risk" in the Selva have a 28.1% actual dropout rate, compared to 38.9% in the Sierra, meaning the same risk score carries different meaning across regions.

Poverty: A monotonic relationship emerges across poverty quintiles—students in poorer quintiles are flagged more frequently and have higher base dropout rates. This alignment between base rates and flagging rates means poverty-based disparities are partially expected, though the magnitude of the FNR gap between the poorest and wealthiest quintiles warrants attention.

Sex: The gender gap is minimal, with an FNR difference of only 0.026 between male and female students. Sex is not a major axis of disparity in this model, consistent with the relatively small gender gap in Peruvian school enrollment at the primary and secondary levels.

Nationality: With only 27 non-Peruvian students in the test set, this dimension is unusable for reliable fairness inference. We report it for completeness but note that any metrics computed on such a small sample are unreliable.

Age: Older students (ages 15–17) are flagged more accurately than younger students (ages 6–11), reflecting both higher base dropout rates among older students and the model's heavy reliance on age as a predictive feature.

6.3 Intersectional Analysis

Table 9. 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.

Table 9 and Figure 6 present the paper's starkest finding. 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, directly demonstrating the intersectionality imperative articulated by Crenshaw [11] and operationalized computationally by Buolamwini and Gebru [8].

The mechanism behind this disparity is interpretable: the model predicts dropout primarily through spatial-structural features—nightlight intensity, district historical dropout rates, census literacy rates—that code indigenous communities as rural. Urban indigenous students "break the spatial profile": they live in urban areas with higher nightlight intensity and lower district-level dropout rates, but face the same educational barriers (language, cultural mismatch, discrimination) as their rural counterparts. The model has no pathway to identify them because the features that capture indigenous disadvantage in rural settings do not activate in urban ones. We note the sample caveat: $n = 89$ for urban other-indigenous students in the test set, which is sufficient for estimating proportions but should be interpreted with appropriate caution.

6.4 SHAP Interpretability

Table 10 and Figures 7–8 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.

Critically, the top 5 SHAP features have zero overlap with the top 5 logistic regression features (which are dominated by indigenous language dummies). This paradigm difference reflects how tree-based models route predictions differently from linear models: where logistic regression must assign large coefficients to identity features to capture group-level differences, LightGBM can achieve

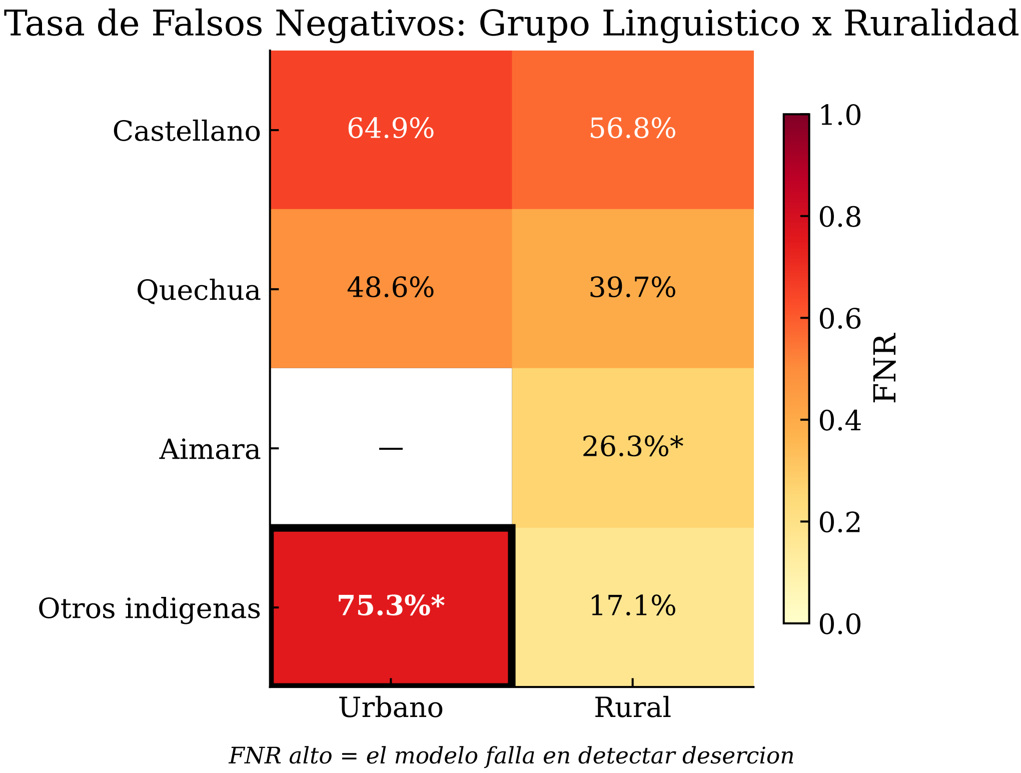

Fig. 6. FNR heatmap by language and rurality intersection. The darkest cell (other indigenous, urban) represents the group most missed by the model.

similar discrimination through the continuous spatial-structural features that correlate with those identity categories. The fairness implications are significant: the model encodes structural inequities without using identity features directly. Removing protected attributes from the feature set would not mitigate the disparities documented above, because the model already operates through proxy features that carry the same information.

7 Discussion

7.1 Summary of Findings

Our proxy equity audit reveals three principal findings about what disparities can emerge from survey-derived dropout prediction in Peru's multilingual, geographically diverse context, each corresponding to a research question.

In response to **RQ1** (what disparities exist in prediction accuracy across demographic groups), we document a surveillance–invisibility axis across language groups: indigenous-language speakers experience low FNR (0.22) but high FPR (0.52), constituting surveillance bias, while Spanish speakers experience high FNR (0.63) but low FPR (0.18), constituting invisibility bias. This systematic redistribution of prediction errors follows from the impossibility result (Section 6) applied to groups with heterogeneous base rates.

Table 10. 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

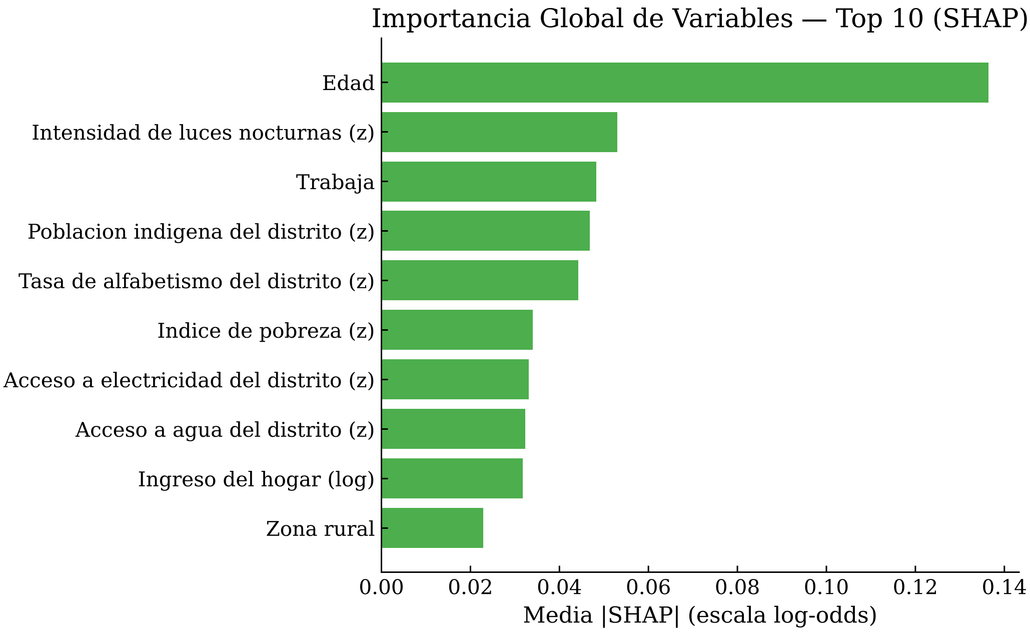

Fig. 7. Mean absolute SHAP values for the top 15 features. Age and spatial-structural features dominate, while identity features (language, sex) have minimal direct importance.

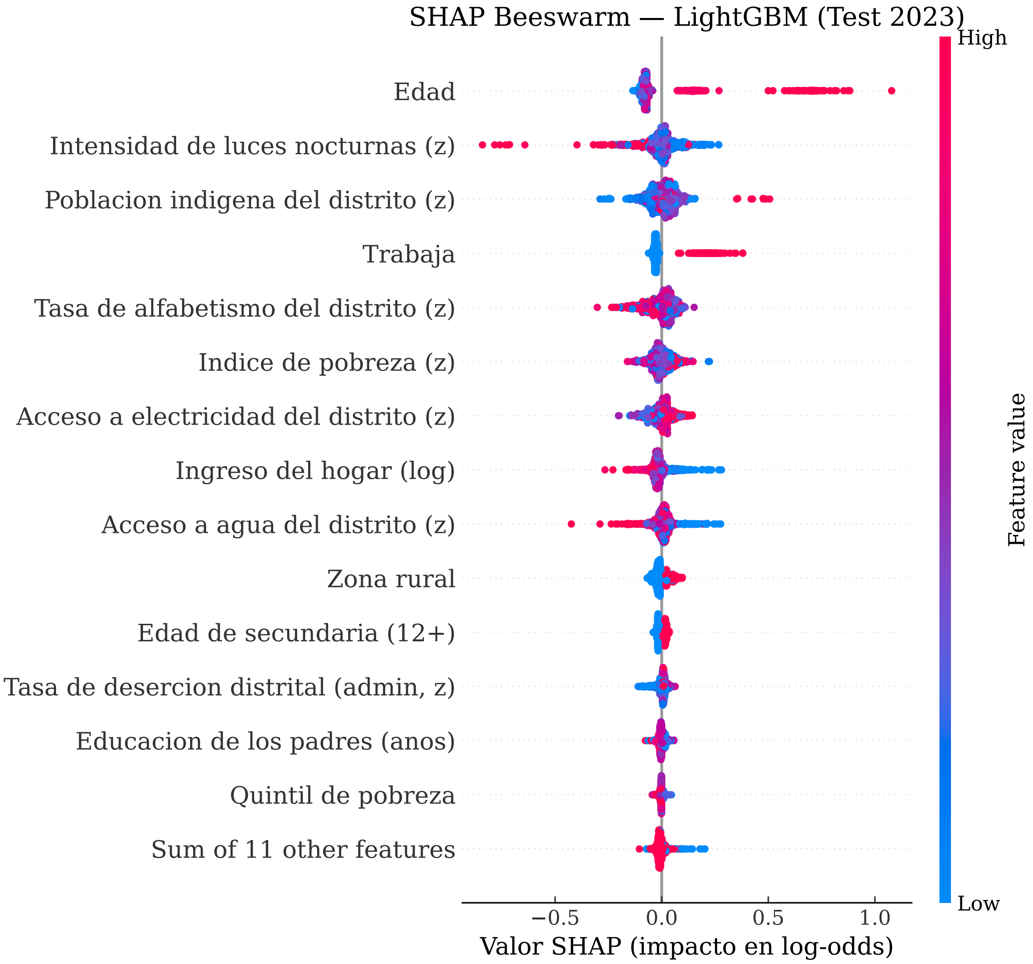

Fig. 8. SHAP beeswarm plot showing feature value distributions and their impact on predictions. Red indicates high feature values; blue indicates low values.

In response to **RQ2** (how the model encodes disparities), SHAP analysis reveals that the model predicts through spatial-structural proxy features—age, nightlight intensity, census literacy rates, poverty index—rather than through identity features directly. Indigenous language indicators, which dominate the logistic regression model, contribute minimally to LightGBM predictions (zero overlap in top-5 features between models). This means the model encodes structural inequities without explicit use of protected attributes.

In response to **RQ3** (how intersections amplify prediction errors), the language × urbanicity intersection reveals urban indigenous students' FNR of 0.753—a disparity completely invisible in single-axis analysis of either language or geography alone. This finding directly operationalizes the intersectionality imperative (Section 6.3) in an educational context.

7.2 The Spatial Proxy Mechanism

SHAP analysis reveals that the model uses geography as a proxy for demographic risk—nightlight intensity, district-level dropout rates, and census literacy rates collectively encode the spatial concentration of disadvantage. This creates systematic blind spots for populations that do not match spatial stereotypes. Urban indigenous students exemplify this failure: they reside in urban areas with favorable spatial indicators yet face educational barriers comparable to their rural counterparts. The model has no pathway to identify their risk because the features that capture indigenous disadvantage in rural settings do not activate in urban ones. This resonates with Perdomo et al.'s [31] argument that structural features predict dropout well—and extends it by showing that reliance on structural features creates predictable fairness failures at demographic intersections.

7.3 Considerations for EWS Operators

Our findings raise several considerations for operators of educational early warning systems:

- Group-specific threshold adjustment could equalize FNR across language groups, reducing invisibility bias for Spanish speakers without necessarily increasing overall error. However, threshold adjustment redistributes errors rather than eliminating them—equalizing FNR would increase FPR for Spanish speakers—and the appropriate trade-off depends on the relative costs of missed dropouts versus false alarms in specific operational contexts.
- Supplementary identification mechanisms for urban indigenous students could address the intersection-level blind spot our analysis reveals. However, designing such mechanisms without creating additional surveillance of already-marginalized communities requires careful consideration of community perspectives and consent [25].
- Regular fairness auditing, conducted across multiple demographic dimensions and their intersections, could detect disparities before they become entrenched. The question of who should conduct such audits—system operators, independent researchers, affected communities, or regulatory bodies—remains open.

7.4 Generalizability

Our findings likely apply to similar EWS systems across Latin America and other developing regions that use spatial features for dropout prediction. The surveillance–invisibility dynamic may emerge wherever prediction models operate on populations with heterogeneous base rates and correlated spatial-demographic structure [1, 35]. Whether the specific intersection-level failures we document (urban indigenous invisibility) generalize depends on the degree to which indigenous populations in other countries exhibit similar rural-urban migration patterns.

8 Limitations

Several limitations qualify the interpretation of our findings.

As established in Section 1, this paper audits a proxy model, not the actual Alerta Escuela system. The feature sets differ substantially (Table 1): ENAHO provides demographic and household variables while SIAGIE contains attendance and grade records. Our findings demonstrate what disparities *can* emerge from survey-derived prediction, not what the deployed system produces.

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 is affected by the COVID-19 pandemic. INEI conducted phone interviews rather than in-person household visits, producing a reduced sample (approximately 13,755 observations versus approximately 25,000 in typical years) with 52% null values in the education attendance variable. While we include 2020 in the training data after dropping null records, this year may not represent the same population as in-person survey years.

Fourth, some intersectional subgroups have small samples: $n = 89$ for urban other-indigenous students (our starkest finding) and $n = 27$ for non-Peruvian nationality (rendering this dimension unusable for reliable inference). While $n = 89$ is sufficient for point estimates of proportions, the associated confidence intervals are wide, and results for this group should be interpreted with appropriate caution.

Fifth, while we incorporate FACTOR07 survey weights throughout training and evaluation, the theoretical properties of survey-weighted gradient boosting are not fully established. MacNell et al. [24] found that ignoring survey weights in gradient boosting can affect both prediction accuracy and feature importance rankings, supporting our decision to incorporate weights, but the formal statistical guarantees of weighted ML estimators under complex survey designs remain an active area of research.

9 Ethical Considerations

Positionality. The author is Peruvian, a computer science self-learner at Universidad de Ingeniería y Tecnología (UTEC), and co-founder of Genera, an educational technology startup. This positionality shapes the work in important ways: as someone who believes AI can address education's one-size-fits-all problem, I undertook this audit precisely because systems deployed without scrutiny risk deepening the inequities I aim to address. I am not from the indigenous communities most affected by the disparities documented here, and this work should be understood as an outsider's technical audit rather than a community-centered assessment. A more complete evaluation would incorporate the perspectives of indigenous educators, families, and students directly affected by early warning systems.

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) reveals a surveillance–invisibility axis in which indigenous-language speakers are over-flagged while the majority of Spanish-speaking dropouts are missed, an intersection-level blind spot for urban indigenous students that single-axis evaluation cannot detect, and a model that encodes these disparities through spatial-structural proxy features rather than identity variables—findings that hold across five model families. As educational early warning systems proliferate globally, routine intersectional fairness auditing is essential for identifying who gets missed.

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, Albert W. Berning, Sujith M. Gowda, Siyuan Zhang, and Amanda Hawn. 2020. Predicting K-12 Dropout. *Journal of Education for Students Placed at Risk (JESPAR)* 25, 1 (2020), 28–54. doi:10.1080/10824669.2019.1670065

[4] 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

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

[6] 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*.

[7] 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

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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

[15] 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/.

[16] 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).

[17] 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.

[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

[21] Sungho Lee and Joo Young Chung. 2019. The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. *Applied Sciences* 9, 15 (2019), 3093. doi:10.3390/app9153093

[22] 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.

[23] 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.

[24] 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

[25] 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

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

[27] 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.

[28] 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.

[29] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting Good Probabilities with Supervised Learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML).* ACM, 625–632. doi:10.1145/1102351.1102430

[30] 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.

[31] 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).*

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

[33] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[34] UNESCO. 2022. *Education in Latin America and the Caribbean: Challenges, Trends and Policies.* Technical Report. UNESCO Regional Office for Education in Latin America and the Caribbean, Santiago, Chile.

[35] 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