# 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 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 [11, 12]. Peru's Encuesta Nacional de Hogares (ENAHO), conducted annually by the Instituto Nacional de Estadística e

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.

Informática (INEI), provides nationally representative data that enables analysis of these disparities [14]. Because the actual SIAGIE administrative data used by Alerta Escuela is not publicly accessible, 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 [16, 31], 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: 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 [7, 10].
- 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.

We emphasize that this is a proxy audit: we have not accessed Alerta Escuela's predictions, training data, or operational feature set, and make no claims about the deployed system's specific fairness properties. Our findings characterize disparities that can emerge from survey-derived dropout prediction in Peru's demographic context.

## 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. [30] 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 [29] 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. [33] 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

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 [25, 26]. 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 presumably 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) [14]. 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

Table 1. ENAHO vs. SIAGIE Feature Availability. SIAGIE columns are inferred from public documentation [25, 26]; 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) |

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.

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 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 9.

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

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] |

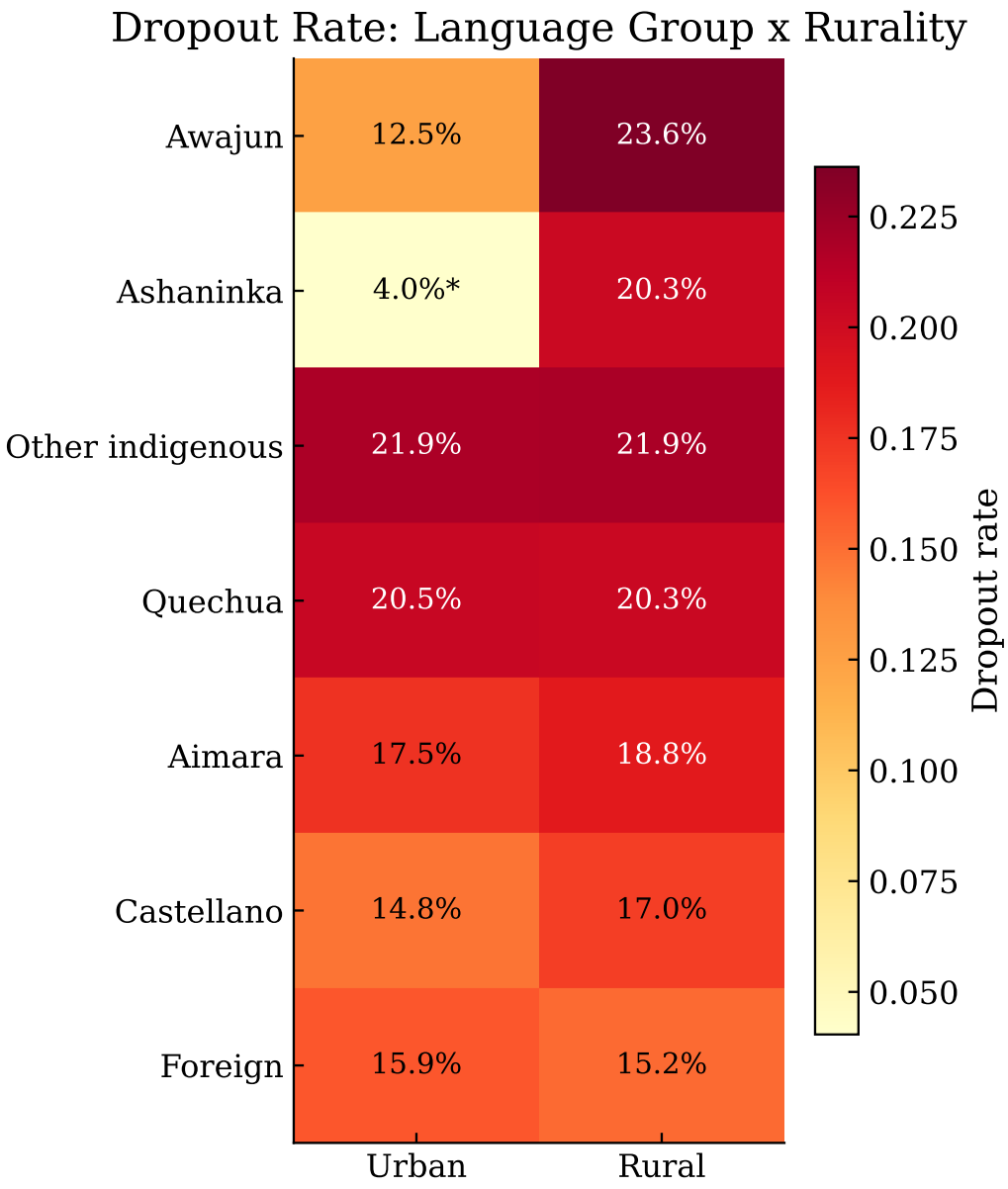


Fig. 1. Dropout rates by language group and rurality. Each cell shows the weighted dropout rate for the intersection of language and urban/rural geography.

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.

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 [23]; 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* [16] serves as the primary predictive model, leveraging gradient boosting's ability to capture nonlinearities and feature interactions. *XGBoost* [8] provides a second gradient boosting implementation for cross-architecture consistency 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 [28]. We apply Platt scaling [31] 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 [5] 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 [3] and aligns with the audit methodology of Saleiro et al. [32].

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 [9] impossibility theorem predicts will arise when base rates differ across groups.

SHAP TreeExplainer [21] 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 |

Near-identical PR-AUC across three tree-based ensembles (Table 5) 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" 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.

Precision-Recall curves and calibration plots appear in Appendix A.

### 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 2). This decile-level concentration of risk confirms that the model's predictions are meaningful, not random.

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 | ↓ |

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 (Figure 2). 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.

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. 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. Section 6 documents those differences across five model families—robustness across architectures provides stronger evidence than any single model's absolute performance.

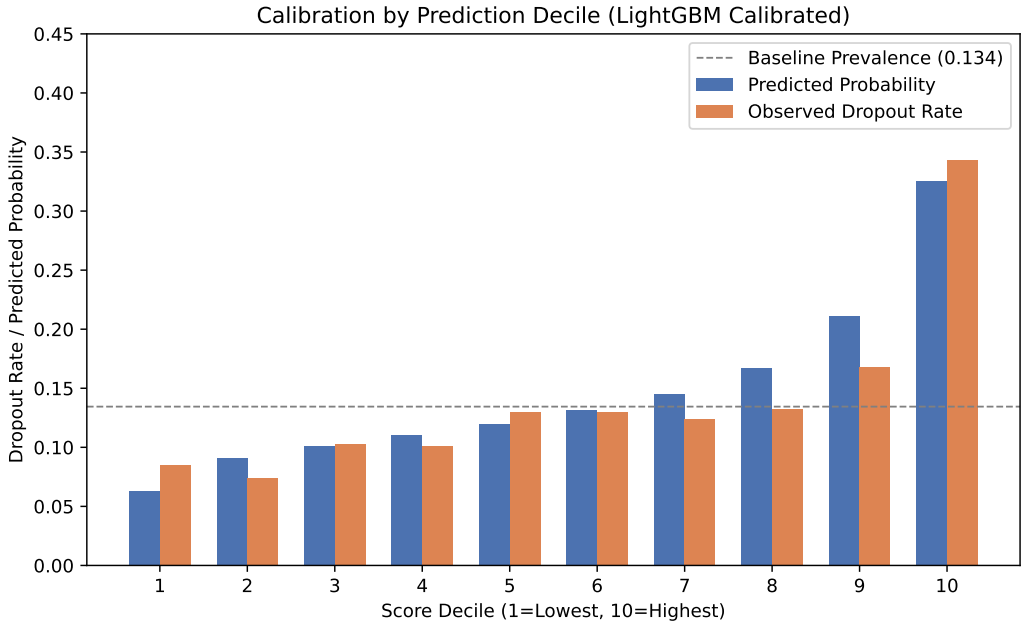


Fig. 2. Calibration by prediction decile for the LightGBM calibrated model. Bars show predicted probability (blue) and observed dropout rate (orange) per decile. Mean absolute calibration error = 0.018, indicating well-calibrated predictions. Baseline dropout prevalence = 0.134 (dashed line).

### 5.2 Cross-Architecture Consistency

Table 7 extends the cross-architecture consistency 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); 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.

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. The MLP Aimara FNR of 0.830 is a small-sample outlier ($n = 76$).

### 5.3 Feature Ablation

Table 8. 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 |

Table 8 tests whether the FNR disparity is driven by spatial features, individual features, or their combination. Castellano speakers have the highest FNR in all three variants, confirming that their invisibility is not an artifact of a particular feature set. The spatial-only model dramatically reduces castellano FNR from 0.633 to 0.317—when the model can only use district-level features, it flags students in disadvantaged districts regardless of language, compressing the language-based FNR gap. Conversely, the individual-only model slightly increases castellano FNR (0.649) while halving indigenous FNR. The full model's combination of spatial and individual features produces the worst outcome for castellano speakers: spatial features provide the primary prediction signal, but individual features allow the model to differentiate within districts, effectively "un-flagging" castellano students whose individual profiles do not match the dropout risk pattern that spatial features establish.

## 6 Fairness Analysis

Table 9. 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 9 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 FNR-FPR relationship is not a model bug but the mathematical consequence of Chouldechova's [9] 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).

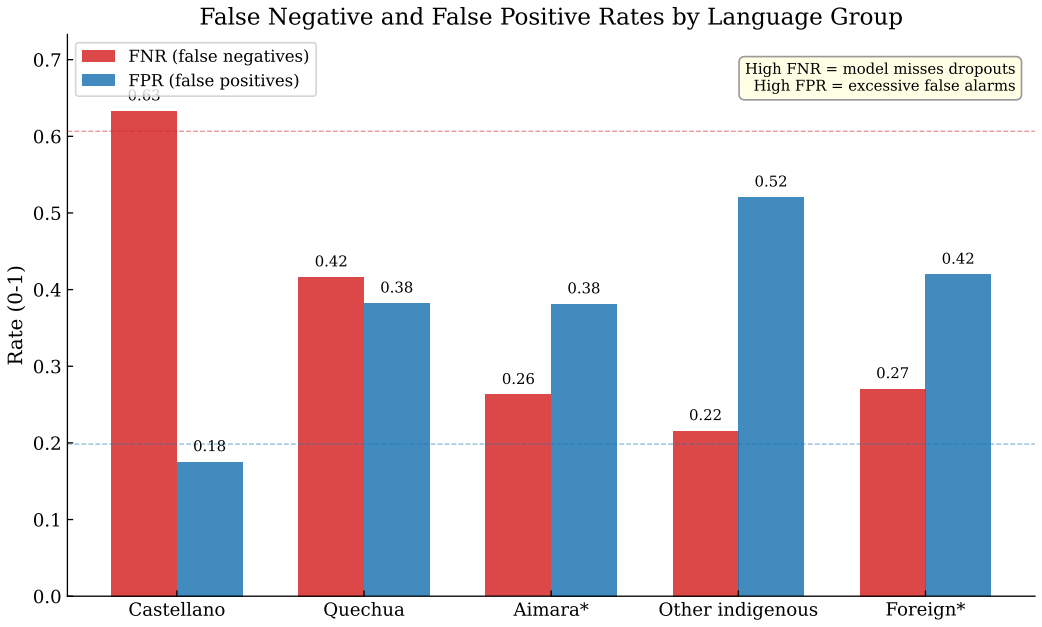


Fig. 3. FNR and FPR by language group. The inverse relationship between FNR and FPR reveals the surveillance–invisibility trade-off.

The inverse relationship between FNR and FPR across language groups (Figure 3) forms a clear axis: as FNR decreases, FPR increases, 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

*Region:* Selva and Sierra show lower FNR than Costa—the model detects rural and remote dropouts more effectively because these students match the spatial profile associated with dropout risk. A calibration gap exists: students predicted as "high risk" in the Selva have a 28.1% actual dropout rate versus 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. Any metrics on this 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 10. 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 10 presents the intersection-level analysis (see also Figure 8 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].

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. Sample caveat: $n = 89$ for urban other-indigenous students in the test set.

Table 11. 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

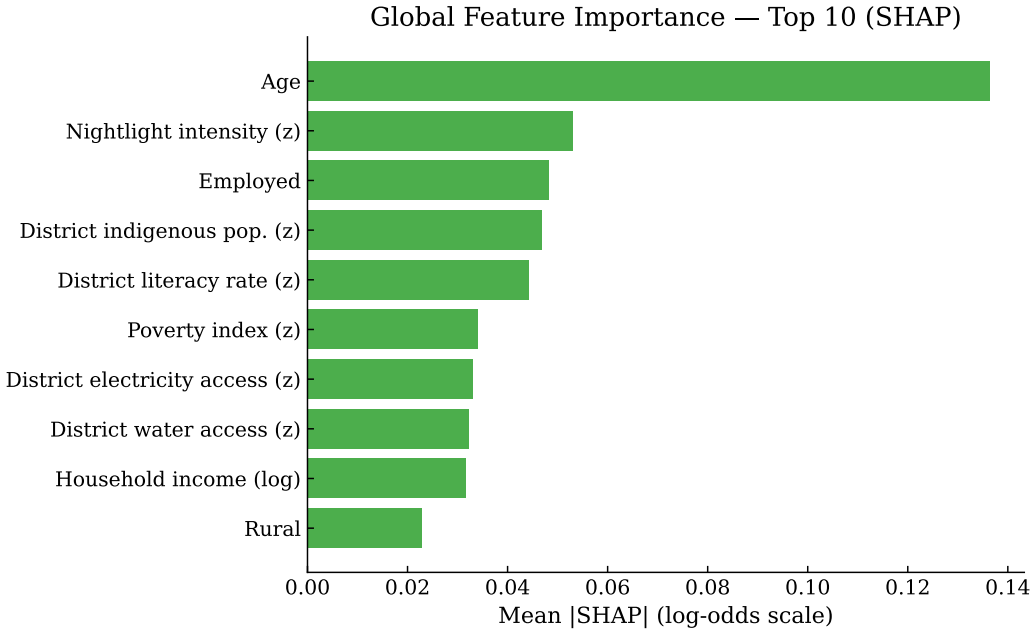


Fig. 4. Mean absolute SHAP values for the top 15 features. Age and spatial-structural features dominate, while identity features (language, sex) have minimal direct importance.

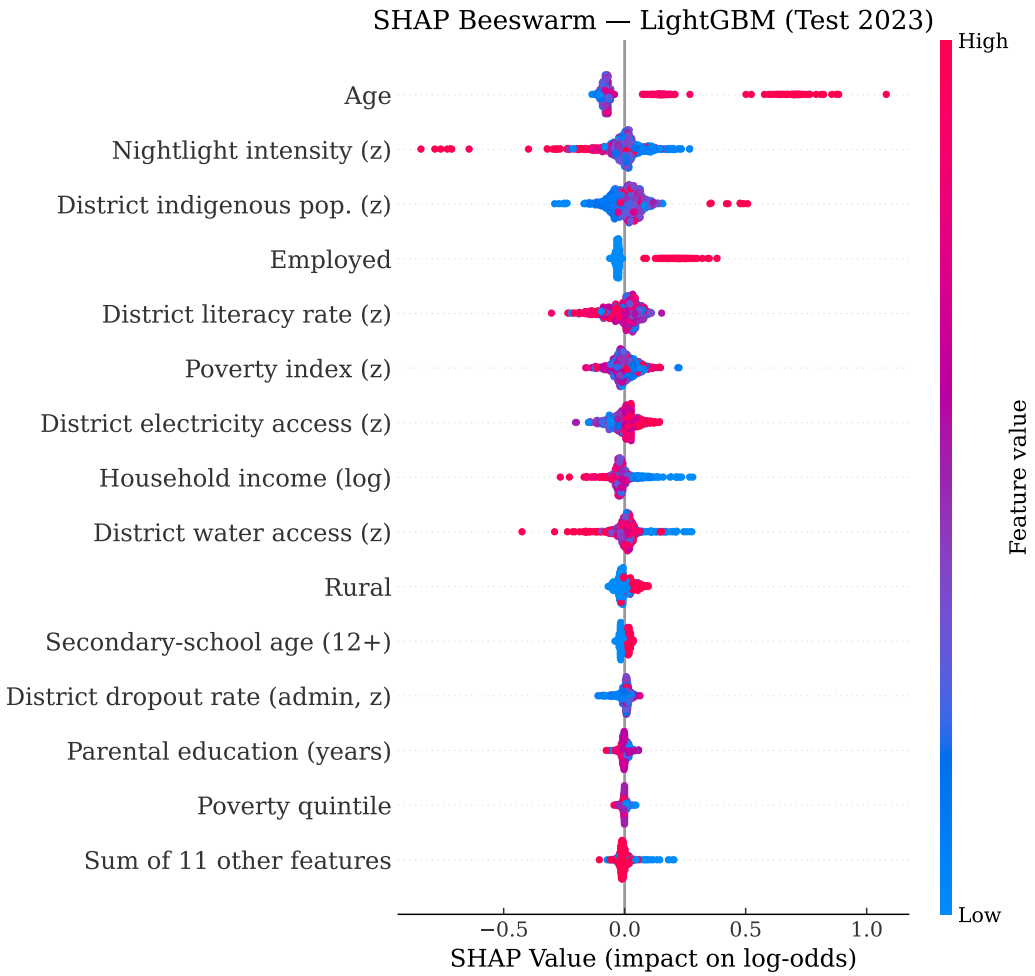


Fig. 5. SHAP beeswarm plot showing feature value distributions and their impact on predictions. Red indicates high feature values; blue indicates low values.

### 6.4 SHAP Interpretability

Table 11 and Figures 4–5 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.

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 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 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 [30] argument that structural features predict dropout well—and extends it by showing that reliance on structural features creates predictable fairness failures at demographic intersections.

### 7.2 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 [24].
- 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.

### 7.3 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, 33]. 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.

### 7.4 Normative Fairness Considerations

Our analysis privileges FNR as the primary fairness metric because a missed dropout represents an irreversible harm: a student who leaves school without intervention faces compounding disadvantage that early warning systems exist to prevent. However, equalizing FNR across language groups would require lowering the classification threshold for Spanish speakers, substantially increasing

their FPR—more non-dropout students flagged, more intervention resources consumed on false alarms. The appropriate trade-off depends on what happens after flagging. If the intervention is low-cost (an automated phone call or teacher notification), elevated FPR is tolerable and FNR equalization is the right objective. If the intervention is high-cost (home visits, social worker assignment), the resource burden of false positives may be prohibitive [24]. We do not resolve this tension—the right criterion depends on operational context we cannot observe as external auditors. What we establish is that the current model implicitly prioritizes low FPR for the majority (Spanish speakers) at the cost of rendering their dropouts invisible [9].

## 8 Limitations

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 in the 2023 test set and $n = 27$ for non-Peruvian nationality (rendering this dimension unusable for reliable inference). Pooling validation and test data yields $n = 167$ urban other-indigenous students with a FNR point estimate of 0.69, but the 95% CI remains wide [0.39, 0.98]. A power analysis quantifies this ceiling: confirming FNR $> 0.50$ at 80% power requires 46 dropout observations, approximately 8 ENAHO survey years at the current rate of ∼6 urban other-indigenous dropouts per year. Detecting the FNR gap between this group and castellano speakers requires 192 dropout observations (∼32 years). This is not merely a sample size limitation but a methodological ceiling: survey-based intersectional fairness auditing fundamentally cannot produce statistically significant results for subgroups contributing fewer than ∼6 positive observations per survey year. This finding argues directly for opening administrative data (SIAGIE), which covers the full student population and could provide the statistical power that survey data cannot.

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. [23] 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) 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

## 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.

[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.*

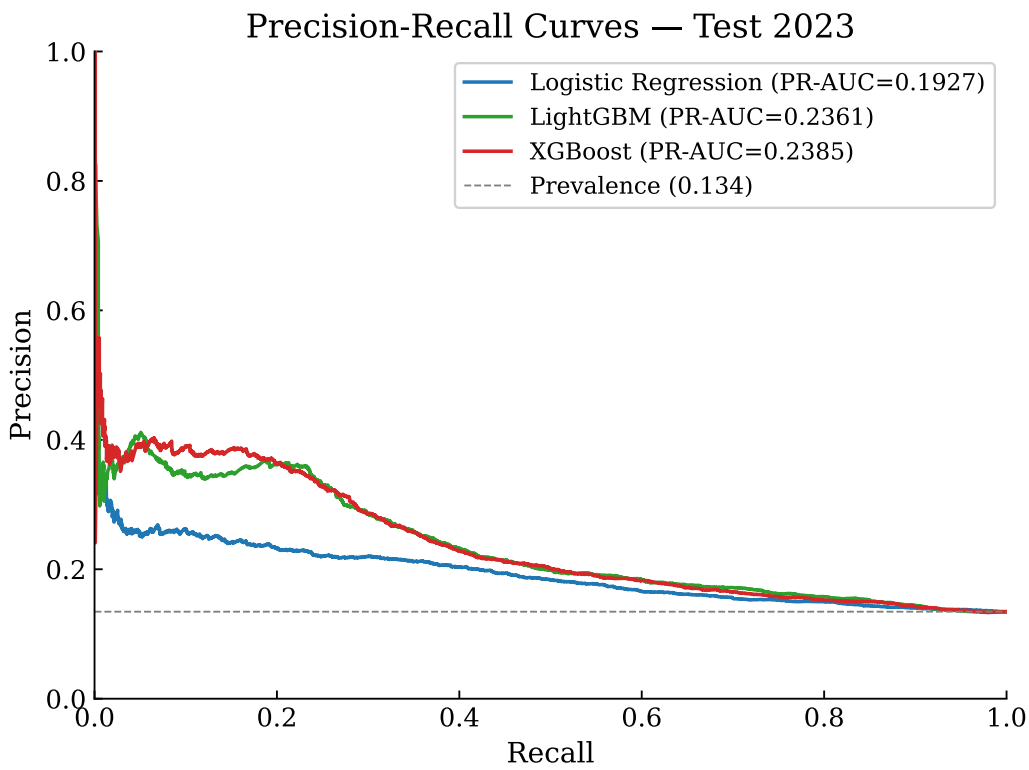


Fig. 6. Precision-Recall curves for three of the five model families on the 2022 validation set. LightGBM and XGBoost curves largely overlap; RF and MLP curves are omitted for visual clarity (see Table 5 for all five).


[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

[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

[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).

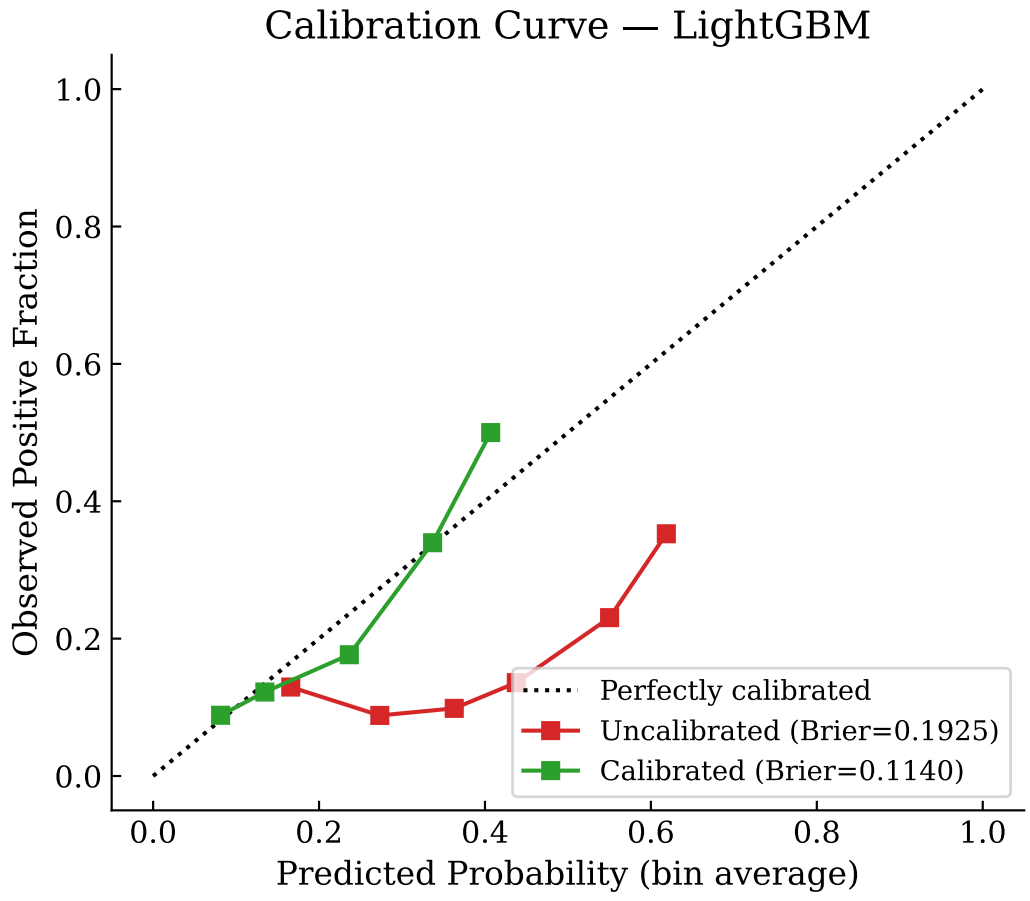


Fig. 7. Calibration plot comparing uncalibrated and Platt-calibrated LightGBM probabilities. Platt scaling reduces test Brier score by 40%.

[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

[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

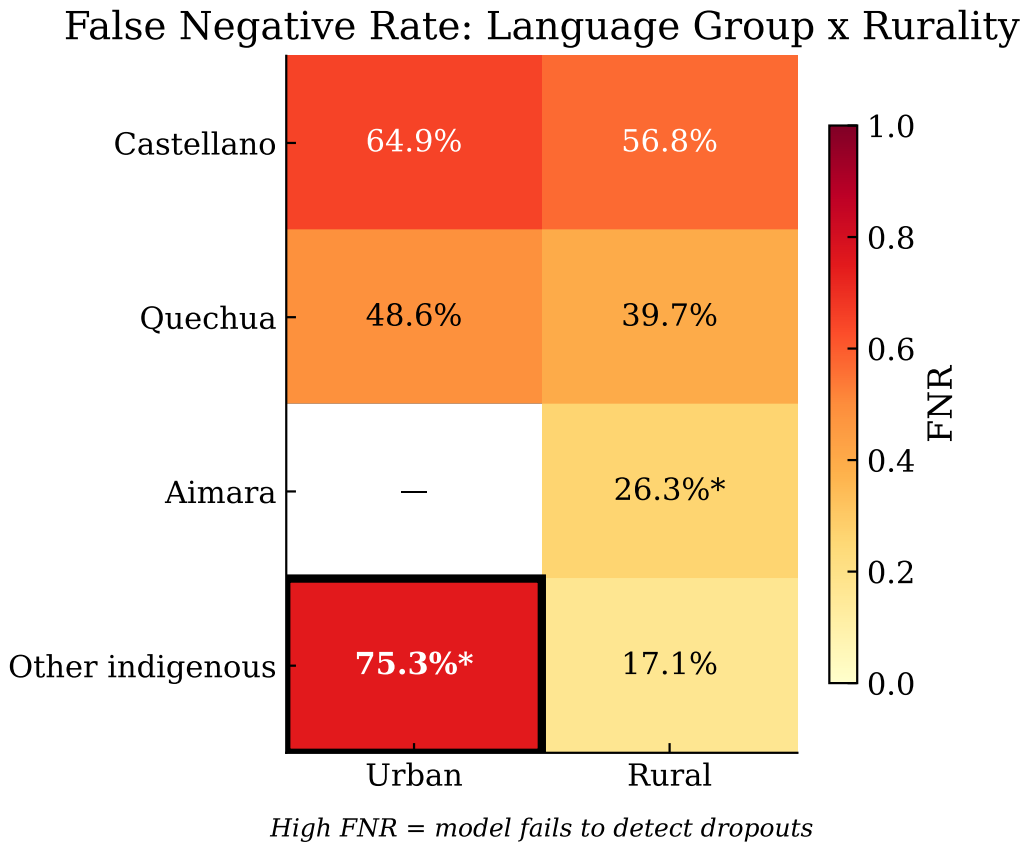


Fig. 8. FNR heatmap by language and rurality intersection. The darkest cell (other indigenous, urban) represents the group most missed by the model.

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.

[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] 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

[29] 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.

[30] 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)*.

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

[32] 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).

[33] 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