Equity Audit of Peru's Alerta Escuela Early Warning System: Who Gets Missed?

[AUTHOR NAME], [Institution], [Country]

Peru's Alerta Escuela early warning system aims to identify students at risk of dropping out, yet its algorithmic fairness properties remain unexamined. Using six years of nationally representative ENAHO survey data (2018–2023, $N = 150{,}135$), we replicate an Alerta Escuela–style dropout prediction model and conduct a comprehensive equity audit across language, geography, poverty, and sex dimensions. Our LightGBM model achieves a calibrated test PR-AUC of 0.236, with a false negative rate (FNR) of 63% for Spanish-speaking students but only 22% for indigenous-language speakers—revealing that the system over-flags indigenous students (surveillance bias) while missing the majority of Spanish-speaking dropouts (invisibility bias). 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, making them the most invisible group. These findings demonstrate that dropout prediction systems can encode systematic inequities even without explicitly discriminatory intent.

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

1 Introduction

Early warning systems (EWS) for student dropout have become central to education policy in Latin America, yet their fairness properties remain largely unaudited [1, 3]. Peru's Alerta Escuela system, operated by the Ministry of Education (MINEDU), uses administrative data to flag students at risk of leaving school [10]. While such systems promise to reduce dropout rates, algorithmic fairness research has shown that predictive models can systematically disadvantage marginalized groups [11].

This paper presents a comprehensive equity audit of an Alerta Escuela–style dropout prediction model, trained on six years of Peru's nationally representative household survey (ENAHO, 2018–2023). We evaluate fairness across language, geography, poverty, and sex dimensions, and use SHAP interpretability analysis to understand how the model makes predictions [9].

2 Related Work

Algorithmic fairness in education has received growing attention. Prior work has examined bias in college admissions [7], automated essay scoring, and course recommendation systems. In the Latin American context, dropout prediction systems are increasingly deployed but rarely audited for equity [12].

The tension between different fairness criteria—equalized odds, predictive parity, and calibration—is well established [3, 4]. Our work applies these frameworks specifically to Peru's educational context, where indigenous language speakers face structural barriers to educational access [5].

3 Data

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) [5]. Our analysis pools six waves (2018–2023) covering school-age children (6–17 years), yielding 150,135 individual-year observations.

Author's Contact Information: [Author Name], [Institution], [City], [Country], [email].

Table 1. Sample Description by Demographic Dimensions

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

Table 1 summarizes the sample across key demographic dimensions. Dropout is defined as a child of school age who was enrolled in the previous year but is not currently attending, following MINEDU's operational definition.

Table 2. Weighted Dropout Rates by Language Group

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

Table 2 reveals substantial disparities in dropout rates across language groups. Indigenous-language speakers face rates 34% higher than Spanish speakers on average.

Table 3 and Figure 1 show how dropout rates vary by region and poverty level, with the Sierra and Selva regions and the poorest quintile exhibiting the highest rates.

4 Methods

We engineer 25 features spanning individual demographics, household characteristics, and district-level spatial indicators from census and administrative data. Models are trained on 2018–2021 data (n = 98,023), validated on 2022 (n = 26,477), and tested on 2023 (n = 25,635), following a temporal split that mirrors real-world deployment.

Table 3. Weighted Dropout Rates by Region and Poverty Quintile

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

We compare three model families: logistic regression (for interpretability), LightGBM [6] (for predictive performance), and XGBoost (for algorithm-independence verification). LightGBM hyperparameters are tuned via Optuna with 100 trials. Platt scaling calibrates the final model's probability estimates. All metrics are computed with ENAHO survey weights (FACTOR07).

Fairness evaluation uses the fairlearn framework [2] to compute false negative rates (FNR), false positive rates (FPR), precision, and PR-AUC across seven demographic dimensions and three intersections. SHAP TreeExplainer [8] provides feature-level interpretability.

5 Results

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

Metric	Logistic Regression	LightGBM	XGBoost
PR-AUC (val)	0.210	0.262	**0.263**
PR-AUC (test)	0.193	0.236	**0.239**
ROC-AUC (val)	0.604	**0.652**	0.648
ROC-AUC (test)	0.598	**0.634**	0.629
F1 (test)	0.270	**0.295**	0.289
Precision (test)	0.194	**0.239**	0.223
Recall (test)	**0.446**	0.386	0.412

Table 4 compares the three model families. LightGBM and XGBoost achieve near-identical validation PR-AUC (0.262 vs. 0.263), confirming algorithm independence of our fairness findings. The calibrated LightGBM model achieves a test PR-AUC of 0.236 with substantially improved Brier score (0.112 vs. 0.186 uncalibrated).

Table 5 shows the logistic regression coefficients. Indigenous language variables dominate the linear model, with the "other indigenous" group having the highest odds ratio (2.20). Figure 2 compares PR curves across models, and Figure 3 demonstrates the calibration improvement.

6 Fairness Analysis

Table 6 reveals a fundamental FNR–FPR trade-off across language groups. The model achieves low FNR for indigenous-language speakers (0.22 for other indigenous) but at the cost of high FPR

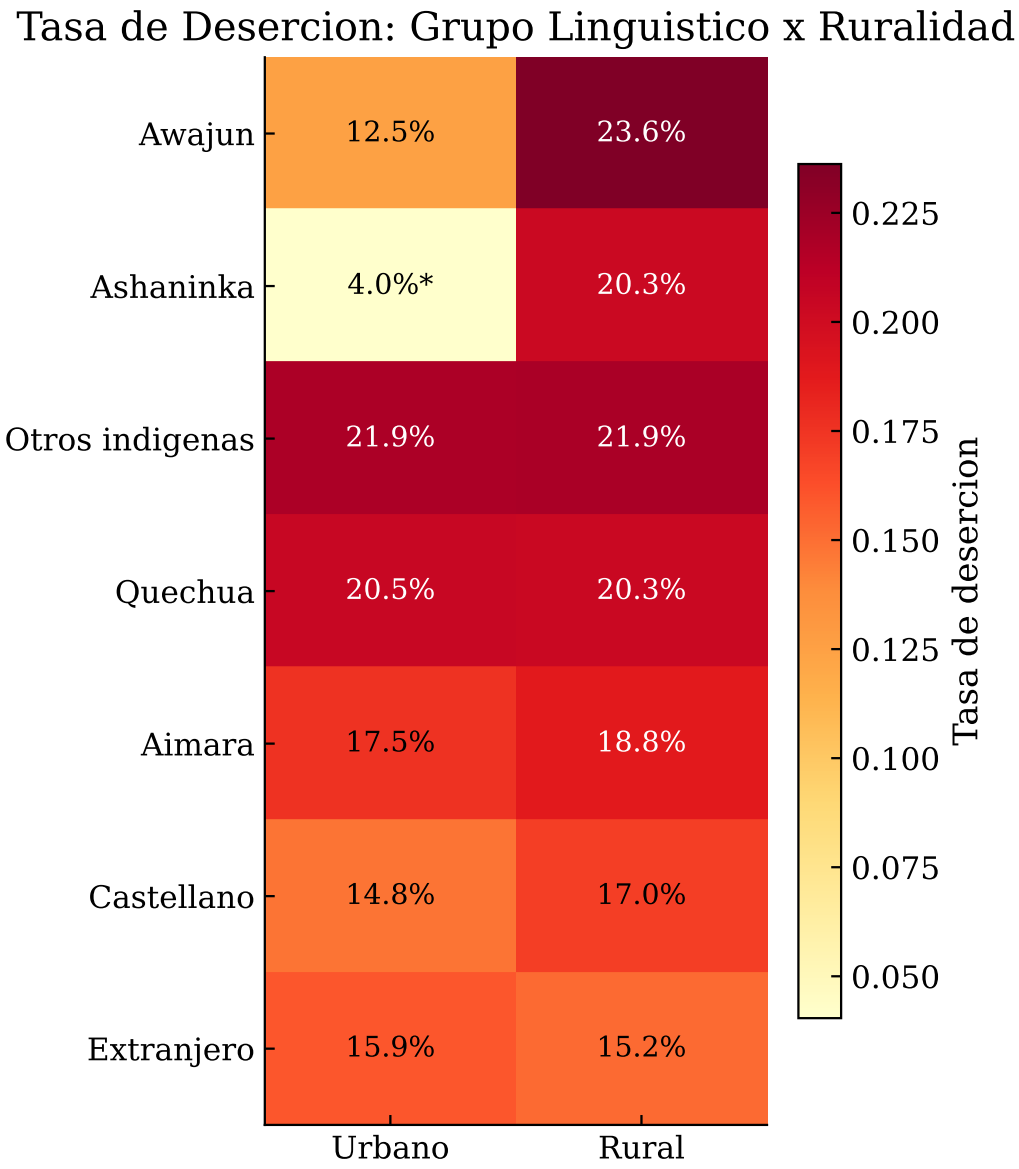

Fig. 1. Dropout rates by language group and rurality. Each cell shows the weighted dropout rate for the intersection of language and urban/rural geography.

(0.52)—a pattern we term "surveillance bias." Conversely, Spanish speakers face high FNR (0.63) with low FPR (0.18)—"invisibility bias" where the majority of actual dropouts are missed.

Table 7 shows the starkest finding: urban indigenous students face an FNR of 0.753—the model misses three-quarters of their dropouts. This intersection group, invisible in both language-only and geography-only analyses, represents the most systematically missed population.

Table 5. 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	↓

Table 6. Fairness Metrics by Language Group

Language Group	*n*	FNR	FPR	Precision	PR-AUC
Otros indígenas	668	0.216	0.521	0.201	0.213
Aimara*	76	0.263	0.381	0.208	0.331
Extranjero*	43	0.271	0.420	0.102	0.425
Quechua	1,624	0.416	0.382	0.221	0.262
Castellano	23,170	**0.633**	0.175	0.243	0.235
Max FNR Gap	—	0.707	—	—	—

Table 8 and Figures 6–7 reveal that the model predicts through spatial-structural features (age, nightlight intensity, working status, census indicators) rather than identity features directly. The top 5 SHAP features have zero overlap with the top 5 logistic regression features, reflecting the paradigm difference between linear and tree-based models.

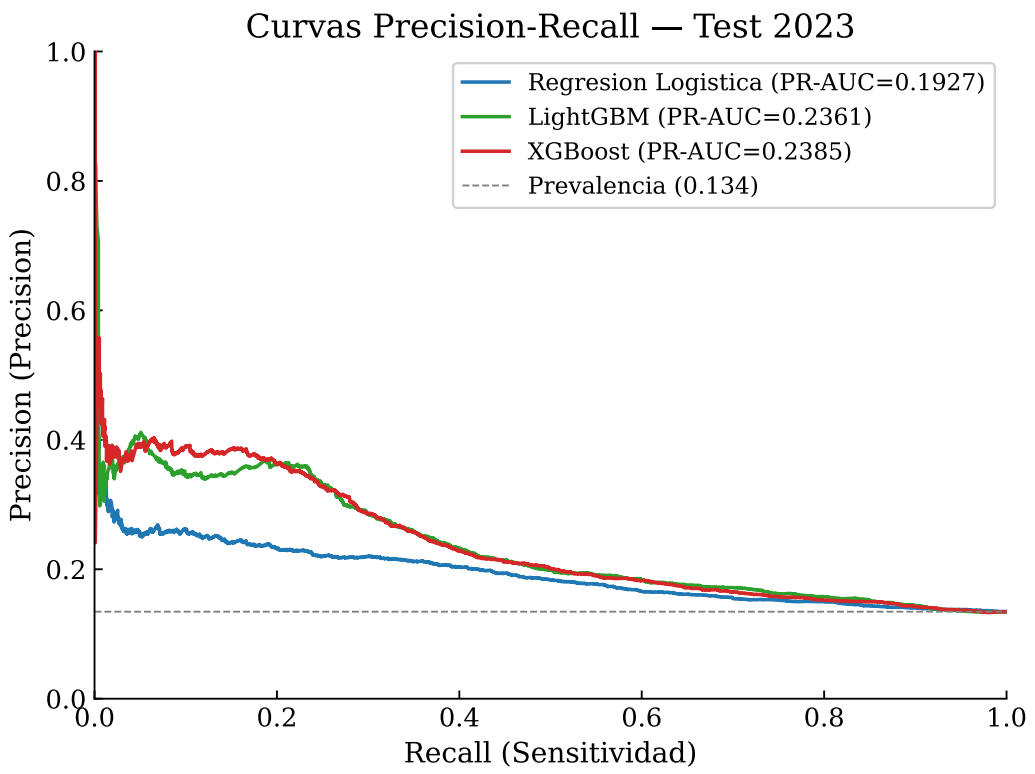

Fig. 2. Precision-Recall curves for all three model families on the 2022 validation set. LightGBM and XGBoost curves largely overlap, confirming algorithm independence.

Table 7. Intersection Analysis: Language × Rurality

Language Group	Urban FNR	Rural FNR	Urban n	Rural n
Otros indígenas	**0.753***	0.171	89	579
Aimara	—*	0.263*	25	51
Quechua	0.486	0.397	234	1,390
Castellano	0.649	0.568	15,598	7,572

* $n < 100$; interpret with caution

7 Discussion

Our equity audit reveals three key findings. First, the FNR–FPR trade-off across language groups represents a systematic redistribution of prediction errors: indigenous students are over-flagged (surveillance) while Spanish-speaking dropouts are missed (invisibility). Second, intersection analysis reveals that urban indigenous students—who defy the spatial profile associated with indigenous languages—are the most invisible group. Third, SHAP analysis shows the model operates through spatial-structural proxies rather than direct identity features, suggesting that even "fair" feature sets can encode structural inequities.

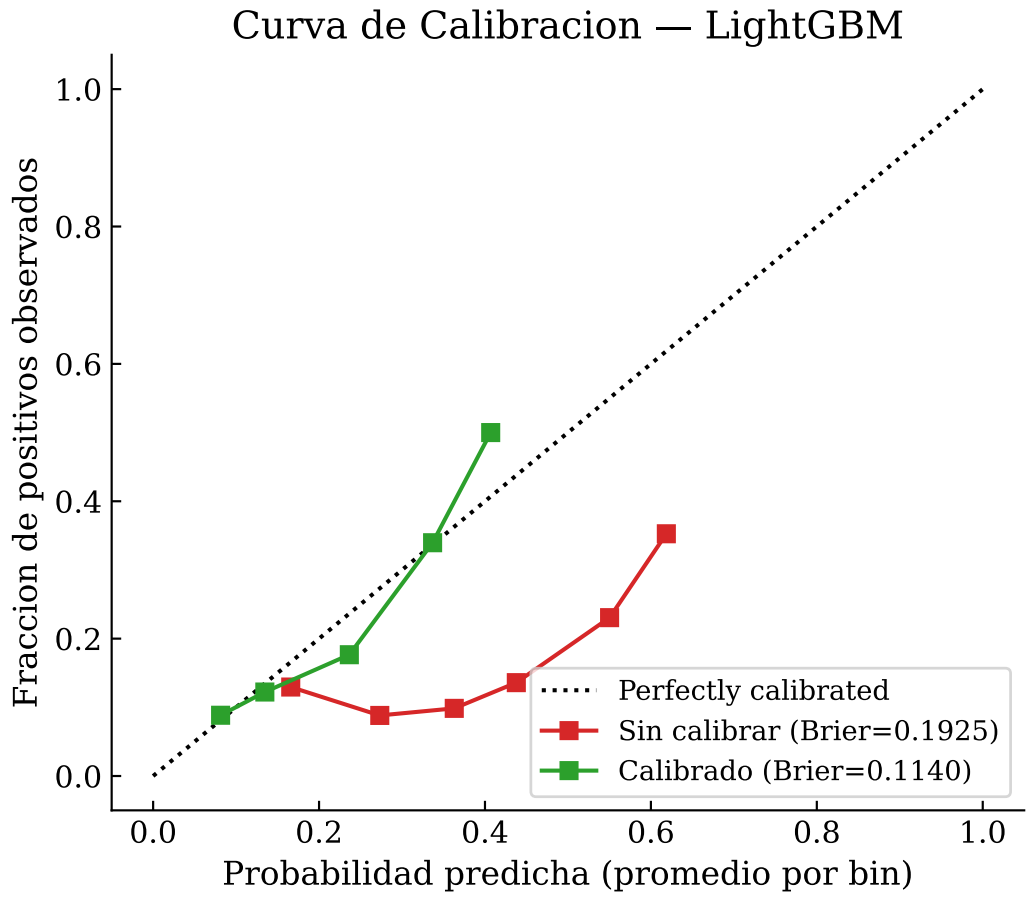

Fig. 3. Calibration plot comparing uncalibrated and Platt-calibrated LightGBM probabilities. Platt scaling reduces Brier score by 38%.

These findings have direct policy implications for Alerta Escuela and similar systems across Latin America. We recommend: (1) group-specific threshold adjustment to equalize FNR across language groups; (2) supplementary urban indigenous identification mechanisms; and (3) regular fairness auditing as a deployment requirement.

8 Conclusion

This paper demonstrates that dropout prediction systems can systematically disadvantage the groups they aim to serve. Our comprehensive equity audit of an Alerta Escuela–style model reveals a surveillance–invisibility axis across language groups and identifies urban indigenous students as the most missed population. These findings underscore the need for mandatory fairness auditing of educational early warning systems.

A Supplementary Tables

Additional disaggregated fairness metrics, regional SHAP decompositions, and model hyperparameter details are available in the supplementary materials.

Table 8. 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

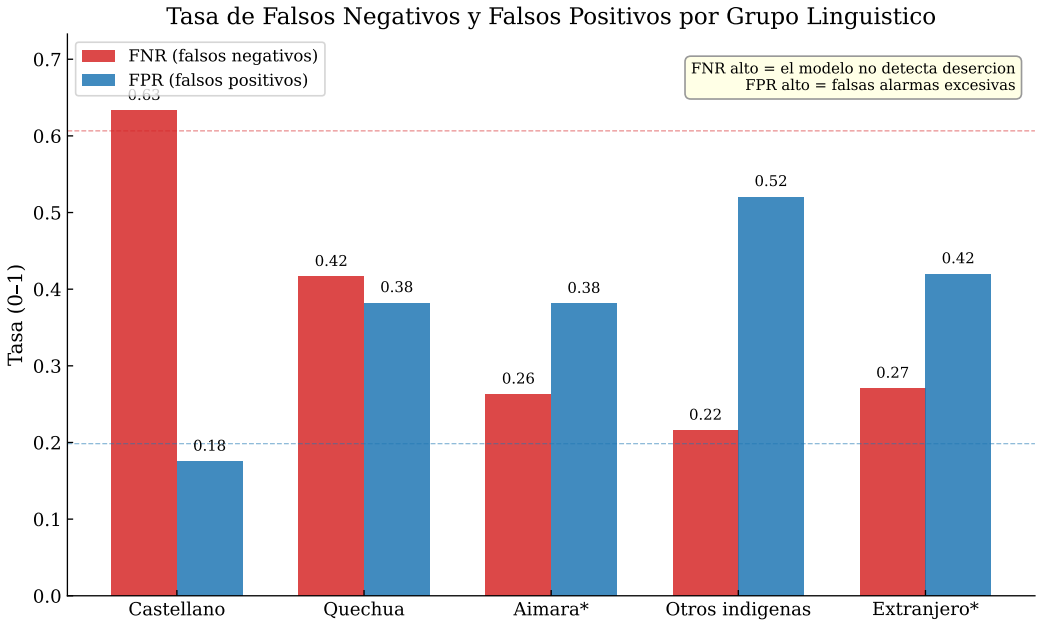

Fig. 4. FNR and FPR by language group. The inverse relationship between FNR and FPR reveals the surveillance–invisibility trade-off.

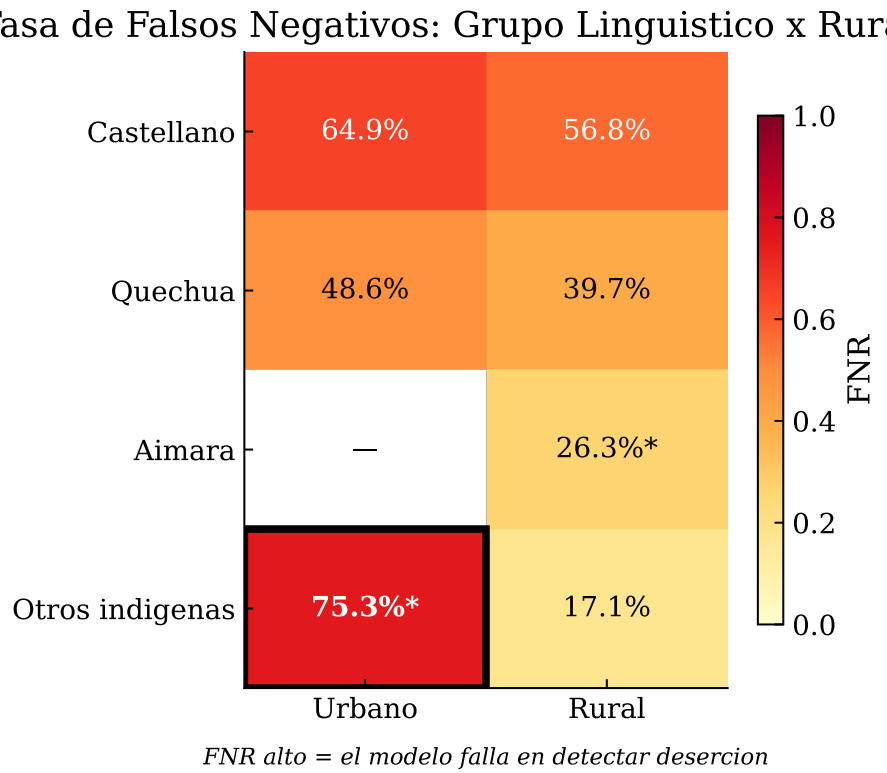

Fig. 5. FNR heatmap by language and rurality intersection. The darkest cell (other indigenous, urban) represents the group most missed by the model.

References

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

[2] 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*.

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

[4] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. In *arXiv preprint arXiv:1808.00023*.

[5] 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/.

[6] 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.

[7] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic Fairness. In *AEA Papers and Proceedings*, Vol. 108. 22–27.

[8] 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.

[9] 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.

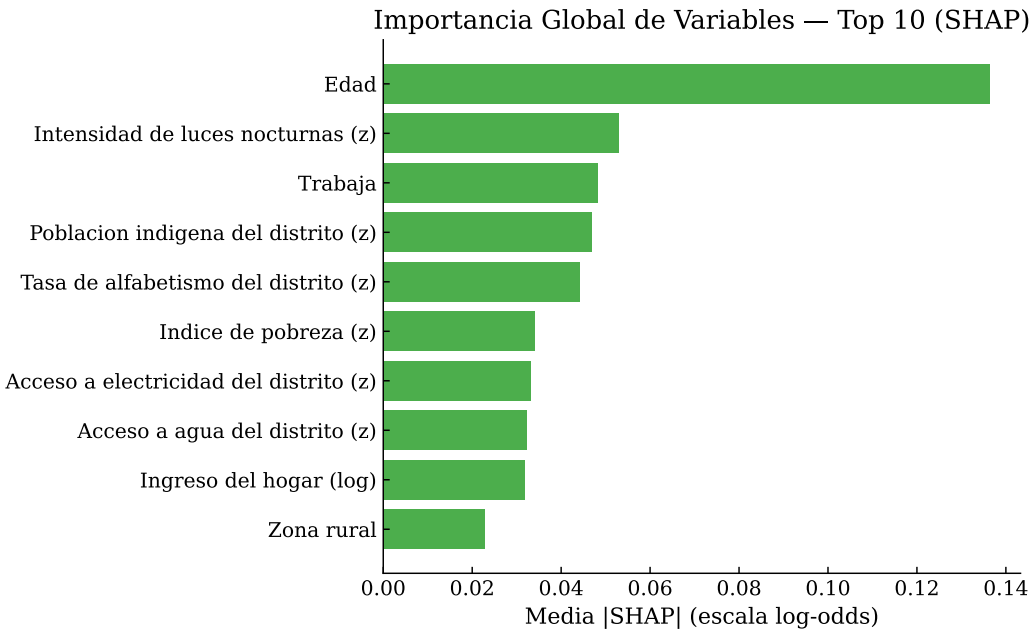

Fig. 6. Mean absolute SHAP values for the top 15 features. Age and spatial-structural features dominate, while identity features (language, sex) have minimal direct importance.

[10] 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.

[11] 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.

[12] 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.

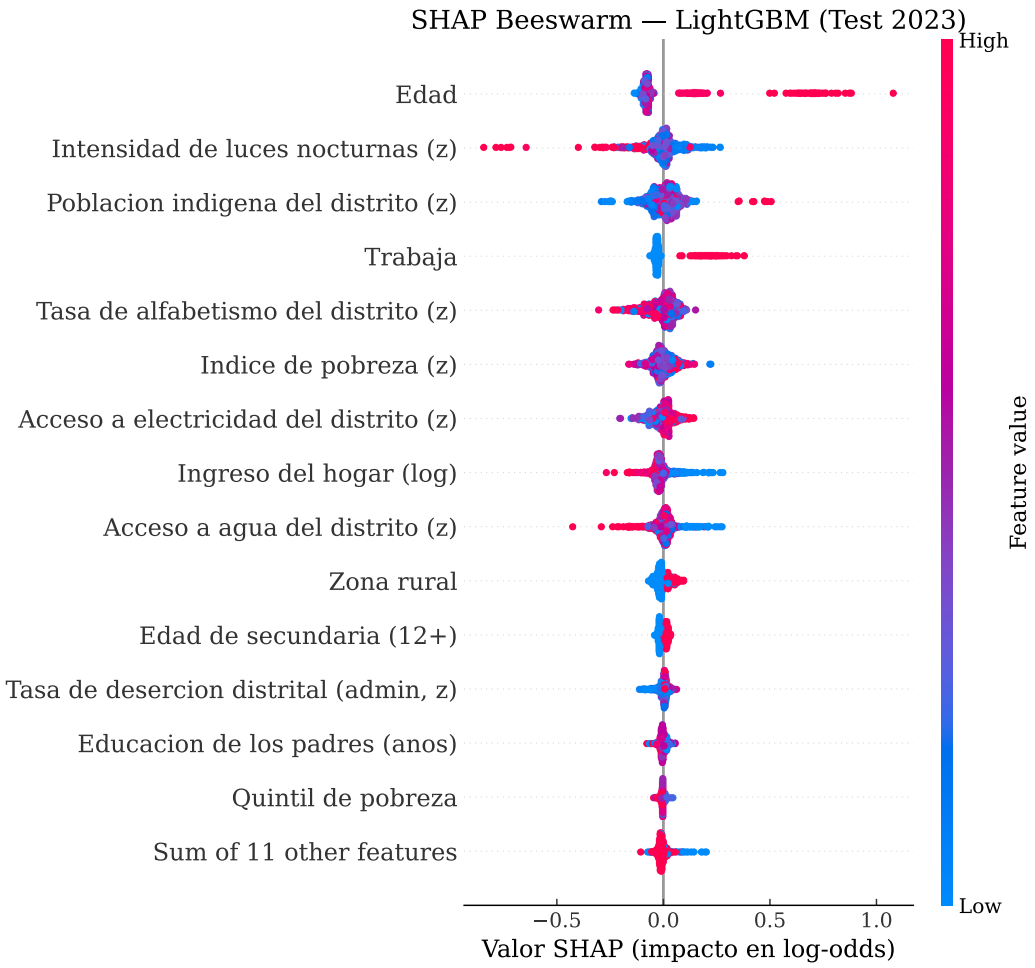

Fig. 7. SHAP beeswarm plot showing feature value distributions and their impact on predictions. Red indicates high feature values; blue indicates low values.