Of Reason, Faith, and Models: An Efficient Machine Learning Approach to Predicting Poverty in Colombia

Juan J. Rojas, Francisco Soler, Jesús Yancy

Universidad de los Andes
Big Data and Machine Learning 2025-2

Best Model Summary XGBoost

-  Kaggle F1 Score of 0.75
- In-sample performance:

  •F1 Score: 0.74 •Precision: 0.71 • Recall: 0.78
- Cross-validation: 5-fold CV, Out-of-Fold threshold optimization.
- Hyperparameters (best tune):
  - *max_depth*: 4 | *min_child_weight*: 10
  - *eta*: 0.05 | *nrounds*: 1000
  - *subsample*: 0.7 | *colsample_bytree*: 0.7
  - *gamma*: 0
- Key predictors: labor-skill & labor-intensity interactions, household structure, income sources and structural determinants (L_p, N_{people}, *P5130*).
- Achieved on a standard laptop (8GB RAM, 4 cores).

Data I Context and Sources

- Source: Gran Encuesta Integrada de Hogares (GEIH) 2018, collected by DANE.
- Dataset:
  Train: 543,109 individuals in 164,960 households
  Test: 219,644 individuals in 66,168 households
- Objective: Predict whether a household is Poor, defined by DANE's regional poverty line (L_p), total household income with imputed rent ($Income_{imp}$), and household size (N_{people}).

$$Poor = \mathbb{I}(\frac{Income_{imp}}{N_{people}} < L_p)$$

- Income: Withheld in the test set to reflect real-world prediction scenarios with limited data.

Structural variables:

- Poverty line (L_p), household size (N_{people}), imputed rent (*P5130*).

Key dimensions captured from individual-level aggregation:

- Labor structure & Household composition: Share of employed, unemployed and inactive relative to household members.
- Human capital:
  - Highest educational level and years of education of the head.
  - Average years of education of employed members.
- Labor intensity and quality:
  - Firm size, tenure and hours worked by head and average among employed.
  - Share and indicator of employed contributing to pension or receiving benefits.
- Non-labor income: Presence of capital or aid income.

Data III Interactions and Imputation Rules

Key interaction variables:

- Education × Share employed: Proxy for skilled labor concentration.
- Hours worked × Share employed: Composite measure of household labor intensity.
- Share employed × Benefits: Captures quality of employment.

Imputation strategy:

- Context-specific imputations based on variable logic.
- Zeros for structurally missing values (e.g., no employees, no rent).
- Categorical NAs assigned to semantically appropriate categories.

Final feature set includes 60 predictors.

Table 1: Summary Characteristics of Poor and Non-Poor Households

	Poor	Non-Poor
Number of observations	33,024 (20%)	131,936 (80%)
Average household members	4.1	3.1
Employed persons per household	1.26	1.56
Average hours worked (head)	28.6	34.5
Imputed rent (COP)	132,000	347,000
Home ownership status	Rented or sublet	Owned, fully paid
Highest education (head)	Primary	University

Notes: Values for numeric variables correspond to sample means. Categorical variables are reported by their sample mode.

- Key insight: Poverty differences emerge from labor intensity, education, and household size, confirming structural economic constraints.
- Precaution: Class imbalance in "pobre", which represents only 20% of the total training dataset.