Problem Set 2: Predicting Poverty

Big Data and Machine Learning for Applied Economics

Prepared by:

Sergio Alejandro Sánchez Martínez (201914432)
Amalia Vargas Guayacán (202424448)
Mariana Villabona Martínez (201816559)

Universidad de los Andes
Department of Economics

Instructor: Ignacio Sarmiento

Submission Date: April 14, 2025

1 Introduction

In Colombia, the efficiency and accuracy of poverty measurement have become increasingly important, as social assistance programs such as Jóvenes en Acción, SISBEN, and Familias en Acción rely on these metrics. Accurate classification is essential to ensure effective targeting and implementation of public policy.

Currently, poverty in Colombia is measured with two main metrics: Multidimensional Poverty Index (MPI) and monetary poverty (DNP, 2017). MPI assesses poverty by evaluating deprivations on five dimensions: education, health, employment, access to public services, and housing conditions. A household is considered multidimensionally poor if it is deprived in at least 33.3% of the weighted indicators (DANE, 2021). The Monetary Poverty Line is determined by the National Administrative Department of Statistics (DANE) and represents the minimum income required per person to meet basic needs. In 2023, this line was set at approximately 435,375 Colombian pesos per month (around USD109). Individuals earning below this threshold are considered to be living in monetary poverty.

To collect data for these measurements, DANE conducts two key surveys: the Gran Encuesta Integrada de Hogares (GEIH) and the Encuesta Nacional de Calidad de Vida (ECV). GEIH is conducted monthly and provides information on labor market statistics, income, and monetary poverty. The ECV is conducted annually and gathers data on various aspects of living conditions, including housing, education, health, and access to services. While these surveys offer comprehensive insights, they are expensive and their periodicity may not capture rapid changes in poverty levels, potentially affecting the timeliness of policy responses.

In order to asses the problem of predicting poverty with high accuracy and at cost-efficient techniques many studies have proposed innovative techniques, such as Astorquiza Bustosa & Muñoz, (2020) that implemented an Progress out of Poverty Index (PPI) as a measurement for current vulnerability and likelyhood of becoming more vulnerable in

Colombia and the Oxford Poverty & Human Development Initiative (OPHI). Others, such as Zixi (2021), Corral et al. (2025), Muñetón et al. (2023) have used Machine Learning approaches, combining spatial analysis and classification models. Previous research points out that employing various ML models including linear regression, decision trees, random forests, gradient boosting, and neural networks and validating through cross-validation and grid search techniques, gradient boosting and random forest as the most accurate model for poverty prediction.

In this study, we predicted poverty status at household level using the following strategies: Logistic regression, Elastic Net, CART, ADA Boost, Grad Boost, XGBoost, Naive Bayes, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Regularized discrimminant analysis and K Nearest Neighbours. To achieve this, we made use of the data from the Colombian National Administrative Department of Statistics (DANE), specifically from the Empalme de las Series de Empleo, Pobreza y Desigualdad project. The dataset was accessed through the following Kaggle competition: Uniandes BDML 202510 PS-2. It includes both individual- and household-level data, divided into training and testing sets. After aggregating the data at the household level, we obtained 164,960 observations for training and 66,168 for testing. However, we do not consider this sample suitable for scaling the models presented in this paper, as it includes unbalanced representation across regions—particularly for cities like Bogotá—and only covers a single year.

On this study, among various algorithms tested, we concluded Gradient Boosting delivered the highest F1 score (0.67), followed by Elastic Net and Logistic Regression (0.62), which effectively balancing precision and recall. Its strong performance reflects its ability to capture complex, non-linear relationships and leverage key variables such as income, education, employment, and household composition. While rebalancing techniques improved classification, high computational costs limited deeper hyperparameter tuning. Future work may benefit from adding richer features and optimizing model training for enhanced performance.

2 Data

2.1 Sources

For this problem set we use data from the Colombian National Administrative Department of Statistics (DANE), collected through the "Empalme de las Series de Empleo, Pobreza y Desigualdad".

We cannot make affirmation on this sample representativity as we do not know to what year the survey belongs. Yet, for today's needs, we do not believe this sample is representative as it only contains approximately 11 million households from the 18 million currently estimated by DANE(2020). Yet, we noticed that test_hogares for household did not contain observations for Bogotá which clearly generates a bias when evaluating the models that predict poverty.

2.2 Missing values and descriptive statistics

The train_dataset provided had 543.109 observations at individual level and 63 variables[1]. We started by exploring the NA values per individual and checking whether missing values made sense or had to be imputed. As the main objective was to measure household level monetary poverty and we did not count with caloric intake variables or nominal monetary income [2] we focused the study on adults and characteristics that could point poverty vulnerability.

We divided data available into three cathegories based on poverty measurement methodology of DANE (2021): Household and Individual Socioeconomic Profile [3] ; Employment-Related Earnings and Additional Income [4] and Non-Wage Compensation, Employment Efforts, and Social Assistance.[5].

[1]The summary of the descriptive statistics can be found on Github file 2- Descriptive Statistics -¿ 1.Exploracion.csv

[2]In Colombia, poverty measurement has utilized both caloric intake assessments and the Orshansky index. Orshansky index estimates poverty by applying a multiplier to the cost of a basic food basket to account for non-food necessities.(DANE,2021)

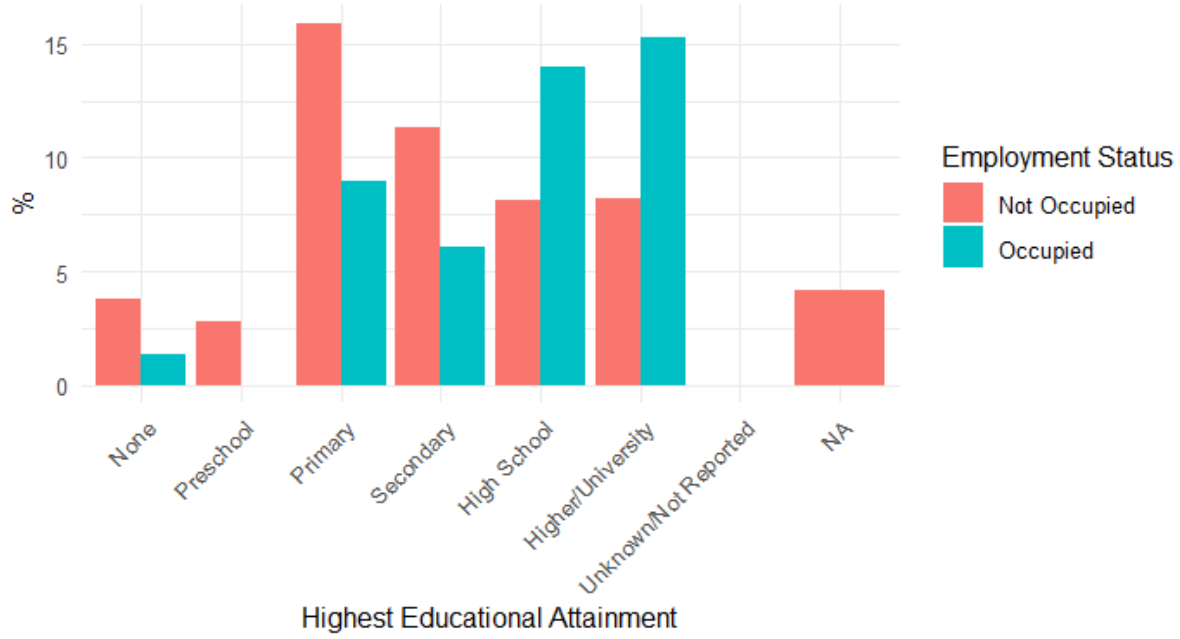
[3]This category includes: Dominio, P5000, P5010, P5090, P5100, P5130, P5140, P6090, P6100, P6210, P6210s1, P6240, Oficio, P6426, P6430, P6510

[4]This category includes: P6510, P6630s1, P6630s2, P6630s3, P6630s4, P6630s6, P6800, P6870, P6920, P7422, P7472, P7495, P7500s2, P7500s3, P7505, P7510s5, P7510s6, P7510s7

[5]This category includes: P6545, P6580, P6585s1, P6585s2, P6585s3, P6585s4, P6590, P6600, P6610,

For the first category, we point out the following points: i) monthly loan repayment (amortization) contained most of missing values, as most of adult individuals were not owners of their dwelling. ii) Only 4% of education level variables were missing on training data set. Thus as studied by Sumarto et al (2006) we considered it as a good candidate for predicting poverty at household level. As we can see on graphic 1 most of individuals that are occupied attain higher levels of education. Reinforcing our belief on the predictive capacity of this variable.

Figure 1: Highest level of education attainment by ocupation type



As per second group missing values in employment-related variables stemmed from individuals outside the working-age population—namely, those under 18 and over 65. Therefore, no imputation was performed for this group. When classifying by job type, we observed that most missing income values corresponded to individuals -the same groups mentioned before. Table 1 presents the distribution of missing values by job type. Notably, 73% of missing data were concentrated among individuals younger than 24 or older than 64.

P6620, P7040, P7045, P7050, P7090, P7110, P7120, P7150, P7160, P7310, P7350, P7510s1, P7510s2, P7510s3

Table 1: Distribution of Missing Job Type by Age Group.

Age group	NA count	TOTAL NA	% of Total
0–12	97,394	295,034	33.01
12–18	50,266	295,034	17.04
18–24	31,466	295,034	10.67
24–65	76,273	295,034	25.85
65+	39,635	295,034	13.43

Finally, the third group was the most interesting when evaluating poverty, as it contained variables related with informality, unemployment, or subsidies. As one of the most important criteria for being selected for a government social subsidies in Colombia is to be classified as a poor household, thus, it's expected the correlation.

2.3 Aggregate variables and data cleaning process

To construct the final dataset for analysis and training classification models, we implemented a data preparation pipeline that merged an individual-level dataset (containing details such as age, education, occupation, and relationship to the head of household) with a household-level dataset featuring variables on housing, location, and demographic indicators. Using the household identifier as the key, we combined these datasets and derived household-level aggregates from the individual data, enriching the household dataset with additional variables to better capture household heterogeneity, including demographic compositions, education levels, employment statuses, and housing conditions, while also performing data cleaning to handle missing values, outliers, and inconsistencies:

Overcrowding (`hacinamiento`) was calculated as the ratio of household members to rooms used for sleeping. Observations with zero or missing room values were reviewed to avoid division errors and no winsorization was applied, as high values reflect real overcrowding conditions. This variable is relevant as insufficient living space and is linked to adverse outcomes like poor health and lower quality of life (Baker, 2008).

Housing cost (`costo_vivienda`) was assigned using rent or imputed rent depending on tenure type. Coded missing values (98, 99) were replaced with the sample median and extreme values were winsorized at the 1st and 99th percentiles to reduce skewness. We believed this variable was relevant as it identifies if households have an excessive housing costs or are living in very low-cost (often substandard) housing depending ont he tenure type.

Dependency ratio (`tasa_dependencia`) was computed as the ratio of non-working to working household members. Missing values, resulting from households with zero employed members, were imputed conservatively using the 95th percentile of the distribution. This variable was proposed following the study of Vijayakumar (2013) who showed proof on high depency rations and poverty on developing countries.

Median education (`median_education`) was aggregated from individual data ignoring missing values. The households with all missing values received the overall sample median. It's already Sumarto et al (2006)

Health vulnerability (`vulnerabilidad`) was set to 1 if all members had poor health or disability.

Informality (`prop_informal`) was computed as the share of informal workers. Households with no employed members had undefined values, which were set to 1, reflecting full informality in the absence of formal work.

Receives bonuses (`recibe_primas`) is an indicator equal to 1 if any household member receives either legal bonuses or additional payments.

Receives subsidies (`recibe_subsidios`) equals 1 if any member received monetary aid from public or private institutions and missing responses were treated conservatively, assuming no subsidies unless at least one is reported.

Non-monetary income (`ing_nomonet`) is set to 1 if any member reports in-kind income sources, such as goods or services received

Receives remittances (`recibe_remesas`) captures whether any household member receives money from relatives or friends and missing values were treated as zeros unless a positive case was found.

Sells durable goods (`des_duradero`) identifies economic distress by checking if a household sold durables and the reason is related to need.

Precarious employment of household head (`empleo_precario_jefe`) equals 1 if the head has a job without monetary remuneration or has a secondary informal job.

Secondary informal job (`segunda_informal`) flags whether any member has a secondary informal job and we treated as 0 when not explicitly stated.

Underemployment intensity (`int_subempleo`) is computed as the share of employed members who want to work more and are available, among all employed. If there are no employed individuals, the ratio is set to 1 to reflect high vulnerability.

Rural: Dummy variable equal to 1 if the household is rural, does not own the dwelling, and is a private household; 0 otherwise.

Labor Vulnerability (`vulnerabilidad_laboral`) is constructed by identifying whether any of five conditions are met for individuals aged 18 to 65: being inactive or unemployed, less than primary education, informal occupation, no pension contribution, and being underemployed. Each condition is encoded as a binary variable, and the final indicator of labor vulnerability is defined as the sum of these five conditions.

2.4 Personal data into Household data

After constructing this variables, we finally constructed a training data set at household level. To achive this, we grouped the training data set at personal level by id, Clase,Dominio,Fex_c,Fex_dpto and Depto in order to guarantee the individuals belonged to the same household. Thus we obtained 164,960 observations on the training set and 45 variables that contained both metrics used to construct the aggregate variables showed on 2.3 and the final results of the aggregated variables. [6].

As per the test data set, as it was alredy set at household level, we only added to

[6]The variables were: id, Clase, Dominio, num_cuartos, num_dormitorios, tipo_posesion, Nper, Npersug, Li, Lp, Fex_c, Depto, Fex_dpto, Pobre, sexo_jefe, edad_jefe, educ_jefe, actividad_jefe, regimen_subsidiado, informal, hacinamiento, costo_vivienda, num_ocupados, num_dependientes, tasa_dependencia, median_education, max_educ_hogar, vulnerabilidad, num_informal, prop_informal, recibe_primas, recibe_subsidios, ing_nomonet, recibe_remesas, des_duradero, empleo_precario_jefe, segunda_informal, int_subempleo, RURAL, CONDICION_LABORAL_FRAGIL, BAJO_NIVEL_EDUCATIVO, OCUPACION_INFORMAL, NO_COTIZA_PENSION, SUBEMPLEADO, VULNERABILIDAD_LABORAL.

the 66,168 original observations and the 16 original columns our estimates for predicting poverty. Thus, the data set also contained 45 variables.

3 Models and Results

To tackle the binary classification task of identifying whether a household is `Pobre` (1) or `No pobre` (0), we first explicitly defined the outcome variable as a factor with meaningful labels, this recoding ensures that the modeling process focuses on directly predicting the condition of poverty, treating it as the positive class of interest. This is particularly relevant in the context of class imbalance, where the `Pobre` category may be underrepresented, and thus metrics such as sensitivity, specificity, and F1-score become more informative than simple accuracy.

3.1 Model Selection and Training

We employed a diverse set of machine learning algorithms to predict household poverty status:

- **Logistic Regression (Logit)**: The first specification used the whole available covariates in the training dataset, excluding only the DANE `DOMINIO` variable. Taking into account the threat of overfitting our model, a second, more parsimonious model was estimated using a subset of 16 carefully selected predictors that are theoretically or empirically linked to poverty, such as education level, housing conditions, household size, and subsidy receipt. This model was used to generate final binary predictions on the test set.

- **Elastic Net**: To allow for variable selection and better handle multicollinearity among predictors, we estimated an Elastic Net regularized logistic regression model using the same set of variables as in the second specification. The Elastic Net approach allowed us to shrink less informative coefficients while maintaining model interpretability and stability.

- **Random Forest (RF)**: Two classification models were developed:
  - *Base Model*: Utilized seven key socioeconomic variables—dependency ratio, household size, median education level, overcrowding, housing cost, vulnerability, and age of the household head.
  - *Extended Model*: Included five additional variables—receipt of subsidies, number of dependents, receipt of remittances, number of bedrooms, and informal sector employment status.
- **Gradient Boosting (GradBoost)**: This model was trained using the same covariate set as the Elastic Net logit to ensure comparability, and standard cross-validation was applied for model evaluation. Details regarding its tuning, performance, and comparison with the other approaches will be further discussed in the Comparative Analysis section.
- **Extreme Gradient Boosting (XGBoost)**: as XGBoost is computationally expensive we used Bayesian Optimization with ParBayesianOptimization package in order to obtain an initial proposal of the best hyperparameters based on an initial model. Thus, two classification models were considered:
  - *Extended Model*: Using all available covariates in the training dataset that were created on 2.3.
  - *Specific Model*: Using only the variables that had more pretictive power, but and slightly mooving the hyperparamethers proposed by the ParBayesianOptimization. Variables used were dependency rate, employment vulnerability, mean of education, overcrowding, vulnerability, rural and subsidies.
- **Naive Bayes**: The Naive Bayes approach makes predictions based on applying Bayes' theorem with strong independence assumptions between features. Unlike more complex models, Naive Bayes offers computational efficiency while maintaining reasonable predictive power for our binary classification problem. Two Naive

Bayes models were developed using different sets of predictors to represent various aspects of household. The first model incorporates twelve variables, including dependency ratio, household size, median education level, overcrowding, housing costs, vulnerability, head of household's age, subsidy receipt status, number of dependents, remittance receipt, number of bedrooms, and informality; and the second model includes thirteen predictors: median education level, housing costs, subsidy receipt status, number of rooms, number of employed household members, vulnerability, healthcare subsidy regime, overcrowding, number of dependents, household size, remittance receipt, housing tenure type, and underemployment. For validation, we employed k-fold cross-validation approach with k=10 to ensure robust evaluation.

- **Linear/Quadratic/Regularized Discriminant Analysis (LDA/QDA/RDA)** We trained three discriminant analysis models to classify poverty status: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Regularized Discriminant Analysis (RDA). LDA and QDA were trained without hyperparameter tuning. RDA was tuned over a grid of regularization parameters: gamma (balancing LDA vs. QDA structure) and lambda (shrinkage intensity of covariance). A grid of five gamma values (0.1–0.9) and two lambda values (0.1 and 0.9) was used to identify the optimal balance between model flexibility and stability. Model performance was evaluated using cross-validated F1 scores. We used only

- **K-Nearest Neighbors (KNN)**: KNN was trained using Using all available covariates in the training dataset that were created on 2.3.

3.2 Hyperparameter Tuning

Each model underwent hyperparameter optimization to enhance predictive performance:

- **Logistic Regression (Logit)**: For the simple logit no hyperparameter tuning was performed.

- **Elastic Net**: the model was tuned via cross-validation over a grid of values for the regularization parameters alpha and lambda, optimizing for the F1-score. We initially explored a broad grid of alpha values from 0 to 1 and lambda values on a log scale from 10^-6 to 10^-2. After identifying a promising region in this space, we refined the grid to focus on alpha between 0.25 and 0.55 and lambda from 10^-6 to 10^-4 resulting in a more precise search for optimal regularization. The final model used the best performing combination (alpha = 0.45, lambda = 1e-04).

- **Random Forest**: A grid search approach was implemented focusing on the `mtry` parameter, exploring values in the range of 2 to 5. This range encompasses the conventional recommendation for classification problems (approximately $\sqrt{p}$) while also considering values smaller and larger to account for potential feature correlations. The number of trees was fixed at 500 to balance model stability and computational efficiency. Performance analysis revealed that:

  - *Base Model*: Optimal performance achieved with `mtry` = 2.
  - *Extended Model*: Optimal performance achieved with `mtry` = 3.

- **Gradient Boosting (GradBoost)**: Details regarding its tuning, performance, and comparison with the other approaches will be further discussed in the Comparative Analysis section as this was the model that out-performed in our study.

- **Extreme Gradient Boosting (XGBoost)**:

  - *Extended Model*: Parameters were obtained by Bayesian Optimization. We assumed a learning rate (eta) of 0.001, The ParBayesianOptimization proposed the following hyperparamethers: max_depth = 15, min_child = 5, subsample = 0.25. Yet, we stablish the subsample at 0.50.
  - *Specific Model*:We did several trials on grids that slightly changed the depth of the network between 5 and 20. We decided the most efficient combination was to keep the depth at 18, we changed the subsample at 0.7, we kept the number of childs at 5 and 500 rounds.

- **Linear/Quadratic/Regularized Discriminant Analysis (LDA/QDA/RDA)**: No hyperparameter tuning was performed.

- **K-Nearest Neighbors (KNN)**: The k-Nearest Neighbors (KNN) model was tuned via cross-validation over a grid of k ranging from 3 to 15 in increments of 2. Model performance was optimized using the area under the ROC curve (AUC), and all predictors were centered and scaled prior to training to ensure equal weighting in distance calculations. The final model selected the optimal value of 7. k=7, which provided the best cross-validated performance within the tested range. Yet it underperdormed.

Table 2: Hyperparameter tuning and model performance

Modelo	α	λ	# Trees	Depth	k (KNN)	m	CV F1
Logit	–	–	–	–	–		0.62
Elastic Net Logit	0.25	1e-6	–	–	–	–	0.66
Random Forest	–	–	500	default	–	2 - 3	0.58
Gradient Boosting	–	0.05	500	6	–	–	0.67
XGBoost	–	0.1	250-500	5-15	–		0.59
Naive Bayes	0	–	–	–	–	–	0.60
LDA	–	–	–	–	–	–	0.47
QDA	–	–	–	–	–	–	0.41
KNN	–	–	–	–	(3,5,7,9,11,13,15)	–	0.53

3.3 Comparative Analysis:

The best performing model in Kaggle was a Gradient Boosting Algorithm (GBM), which achieved a public leaderboard score of 0.6742, outperforming all other submissions from the team (see Table X). This model was trained using the *caret* package in R, with an extensive hyperparameter tuning process conducted via 5-fold cross-validation. The evaluation metric used to optimize the model was the F1 score, which is particularly suitable for imbalanced classification problems.

Table 3: Kaggle competition results by model

Algorithm	**Rebalance**	**Public Score**	**Private Score**
GBoost	✓	0.6742	0.6739
Elastic Net	✓	0.6678	0.6740
GBoost	✗	0.6372	0.6304
Logistic Regression	✗	0.6281	0.6299
Naive Bayes	✗	0.6096	0.6164
Naive Bayes	✗	0.6060	0.6156
Naive Bayes	✗	0.6060	0.6156
Random Forest	✗	0.5851	0.5869
Random Forest	✗	0.5835	0.5948
XGBoost	✗	0.5800	0.5805

The training process involved a comprehensive grid search over several key Gradient Boosting hyperparameters, including: the number of trees, interaction depth (interaction.depth), learning rate, and minimum number of observations in terminal nodes. This tuning was performed in stages, gradually increasing the model's complexity and narrowing down promising parameter ranges. The final grid included learning rates between 0.04 and 0.06, tree depths between 5 and 7, and up to 600 trees, allowing for fine-grained control over model performance.

Model selection was handled by the *train* function in *caret*, with training control options configured to save predictions, compute multiple metrics, and select the best-performing model based on the F1 score. Parallel processing was used to speed up training.

The superior performance of the tuned GBM can be attributed to:

i. its ability to model non-linear and complex interactions.

ii. careful hyperparameter tuning.

iii. the use of threshold calibration to enhance predictive power

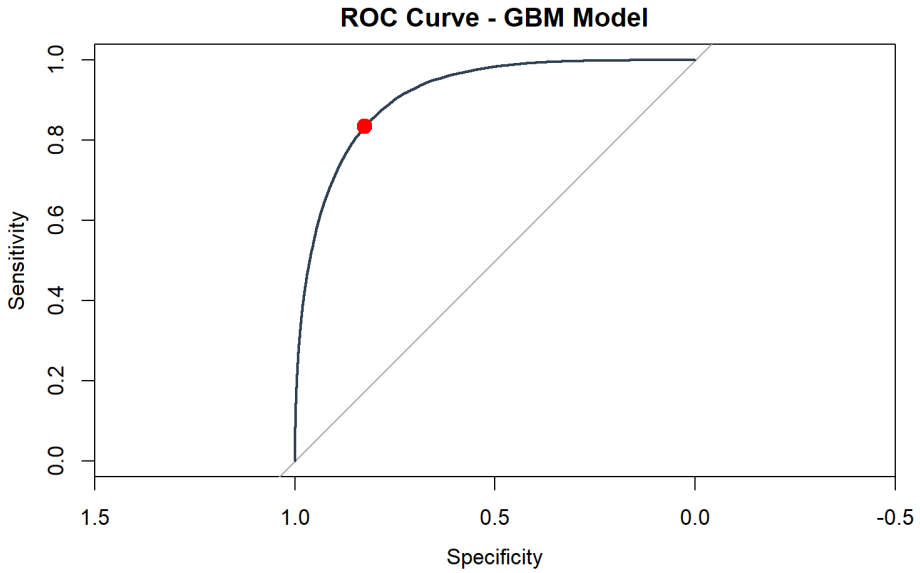
Simpler models like Naive Bayes or classic logistic regression failed to capture intricate relationships in the data, while Random Forest was not optimized or calibrated for the task.

Given the imbalance in poverty (where the `"Pobre"` class represents around 20.7% the training sample) a simple default threshold of 0.5 for classification was not optimal. Instead of applying over- or under-sampling techniques, we adopted a probability threshold adjustment strategy based on the ROC curve.

After training the GBM model, we computed predicted probabilities for each class and used the ROC curve to identify the threshold that maximizes the trade-off between sensitivity (true positive rate) and specificity (true negative rate). Specifically, we selected the point on the curve closest to the top-left corner, which is often referred to as the "best" threshold in ROC analysis.

This threshold, which was 0.3131, improved the recall for the minority class `Pobre`without excessively sacrificing precision. This adjustment helped optimize the F1 Score, which balances both precision and recall.

Below, we include the ROC curve used for this analysis:



This strategy allowed the model to better classify the target and permits to highlight that models without threshold adjustment showed a tendency to underpredict this class, leading to lower F1 scores.

3.4 Feature Importance:

The Gradient Boosting model also provides insights into the relative importance of each predictor variable, based on the total gain across all trees. The most influential variable was Health Vulnerability, with an importance score of 100. Other key predictors included Dependency ratio (55.9), Overcrowding (53.0), and Housing cost (30.6). These variables likely capture structural aspects of the composition of the household and the living conditions that are strongly associated with poverty.

4 Conclusions

In this project, we aimed to clearly and precisely answer the following question: How can predictive models best identify poverty status at the household level? To do so, we trained and compared a wide range of supervised learning models, including logistic regression, elastic net logit, Random Forest, Gradient Boosting, XGBoost, Naive Bayes, LDA/QDA/RDA, and KNN, leveraging both household-level and individual-level information.

Several models showed competitive performance, but the Gradient Boosting Algorithm emerged as the best-performing model in terms of predictive power, achieving the highest F1 score. This suggests that the model was particularly effective at balancing precision and recall—an essential consideration in the context of poverty classification, where both false positives and false negatives can have significant implications.

The superior performance of GradBoost likely stems from its ability to capture complex non-linear interactions and variable importance hierarchies, which simpler models may miss. The best-performing model also made effective use of features from both household and individual levels. In particular, variables such as income level, educational attainment, number of dependents, employment status, and household composition played a crucial role in enhancing predictive accuracy.

A key insight from this work is that rebalancing strategies (such as adjusting the classification cutoff) can meaningfully improve model performance by addressing class

imbalance. However, we also observed that the computational cost associated with training complex models like Random Forests and Boosting methods is high. This makes it challenging to explore larger or more refined hyperparameter grids, and limits our ability to maintain full control over the training process in a reasonable amount of time.

Looking ahead, performance could likely be improved by incorporating new features that capture more nuanced household dynamics or contextual information. While the current models already leverage high-predictive-value variables, expanding the feature set and continuing to refine hyperparameters could further boost the F1 score and overall classification effectiveness.

5 Additional Guidelines

Details on submitting the document in Bloque Neón, the GitHub repository, and code structure are provided to ensure reproducibility and clarity.

References

[1] Alkire, S., Santos, M. E. (2014). Measuring acute poverty in the developing world: Robustness and scope of the multidimensional poverty index. World Development, 59, 251-274.

[2] Baker, J. L. (2008). Urban poverty: A global view. World Bank

[3] Astorquiza Bustosa, B. A., Ospina Muñoz, M. C. (2020). ¿ Menos pobres más vulnerables? Una medición alternativa de la pobreza basada en el progress out of poverty index. Desarrollo y Sociedad, (86), 13-42.

[4] Corral, P., Henderson, H., Segovia, S. (2025). Poverty mapping in the age of machine learning. Journal of Development Economics, 172, 103377.

[5] Departamento Administrativo Nacional de Estadística (DANE). (2020). Proyecciones de viviendas y hogares con base en el Censo Nacional de Población y Vivienda

2018 y censos anteriores. https://www.dane.gov.co/ (replace with actual URL if available)

[6] Muñetón-Santa, G., Manrique-Ruiz, L. C. (2023). Predicting multidimensional poverty with machine learning algorithms: an open data source approach using spatial data. Social Sciences, 12(5), 296.

[7] Sumarto, S., Suryadarma, D., Suryahadi, A. (2007). Predicting consumption poverty using non-consumption indicators: Experiments using Indonesian data. Social Indicators Research, 81(3), 543-578.

[8] Vijayakumar, S. (2013). An Empirical Study on the Nexus of Poverty, GDP Growth, Dependency Ratio and Employ ment in Developing Countries.