A Machine Learning Approach to Analyze Income Determinants in the Mexican Labor Market: The Role of Education (2000–2020)

École Polytechnique, ENSAE, Télécom Paris

Anahi Reyes Miguel[1]

April 2025

1. Introduction

Between the 1980s and 1990s, Mexico experienced a structural transformation from a protected, public-sector-led economy to a globally integrated, private-sector-driven one. Although this shift brought economic stability and growth, income inequality remains high by international standards. As of 2022, Mexico had one of the highest levels of income inequality in the world Chancel et al. (2022). It also recorded the lowest tax-to-GDP ratio among OECD countries, and one of the lowest in Latin America and the Caribbean OECD and BID (2024). This persistent inequality is particularly striking given the country's rapid progress in educational attainment—both in terms of coverage and distribution. Education is typically regarded as a powerful equalizing force; however, in Mexico, as in other developing and developed countries, the expansion of education has not led to a corresponding reduction in income inequality Lopez-Acevedo (2006).

The relationship between education and income has long been a central theme in labor economics (Becker, 1964; Mincer, 1974). While many studies have confirmed that educational attainment correlates with higher earnings, this association can be influenced by labor market structures and social inequalities. In the case of Mexico, income distribution is shaped not only by education but also by informality, regional disparities (Hausmann et al., 2021), and occupational mismatches Quinn and Rubb (2006).

Recent advances in machine learning (ML) provide powerful tools to analyze income dynamics by capturing non-linearities, interaction effects, and temporal changes in a flexible and data-driven way. Unlike traditional econometric approaches—which typically

1. anahi.reyes-miguel@polytechnique.edu

rely on strong assumptions, fixed model structures, and a single estimation based on theory—ML treats empirical analysis as an algorithmic process that systematically explores and compares multiple models. A key feature of this process is tuning, or model selection, which is integrated directly into the algorithm and driven by the data rather than predefined assumptions. This makes ML especially valuable when the functional form of relationships is unknown or complex Athey (2019).

Several recent studies support the use of ML to study income determinants. Matkowski (2021) compares traditional and ML-based income prediction models using U.S. Census data. His work demonstrates that Gradient Boosting and Random Forest outperform OLS, with education consistently emerging as a key predictor—even in the presence of high-dimensional features. Herrera et al. (2023) use XGBoost and neural networks to show how education and digital access jointly shape income distribution in Brazil using survey data. They rely on SHAP values to reveal complex feature interactions, highlighting the importance of interpretability in ML-driven policy research. And In the Mexican context, Gomez-Cravioto et al. (2022) predict alumni income in Mexico using a combination of econometric and ML models using survey data. Education-related features such as GPA and field of study rank among the most important predictors, with SHAP values used to interpret results. Despite its focus on a specific (elite) university population, the study underscores the relevance of ML for understanding wage trajectories.

Building on these insights, this study applies machine learning (ML) models to harmonized Mexican census data from 2000 to 2020 to assess the evolving role of education in income prediction, with a focus on interpretability. The central research question is:

> *Has the importance of education in predicting income increased over time (2000–2020), relative to other factors such as age, gender, occupation, or local labor market conditions in Mexico?*

Using models such as Linear Regression, Lasso, Ridge, Random Forest, and Gradient Boosting, this project evaluates the predictive contribution of education alongside other sociodemographic and labor-related variables. SHAP values are used to enhance model transparency and trace how the relevance of education has changed over the past two decades.

2. Data

2.1 Dataset Description

This study draws on harmonized census microdata compiled by Banco de México, which includes individual-level information from the extended questionnaires of the 2000, 2010, and 2020 Population Censuses, as well as the 2015 Intercensal Count. It includes sociodemographic and labor-related variables, allowing for consistent comparison over time. It offers complete national coverage and municipal-level representativeness. Due to computational and processing limitations, the analysis was conducted on a simple random sample of 72,000 observations drawn from the full dataset. [1]

The target variable for prediction is the **monthly labor income**, denoted as `INGRESO`. The analysis focuses on individual-level characteristics and contextual labor market factors that influence income levels across the Mexican population between 2000 and 2020. The main explanatory variables reflect commonly studied determinants of earnings in the literature (Hausmann et al., 2021), which are the following:

- **Educational attainment** (`ESCOLARIDAD_ACUMULADA`): Accumulated years of formal schooling.
- **Age** (`EDAD`): Continuous variable representing the individual's age in years.
- **Gender** (`LLAVE_SEXO`): Binary indicator for male (1) and female (2).
- **Ethnic identification** (`LLAVE_PERTEINDIGENA`, `LLAVE_AFRODES`): Categorical variables indicating self-reported Indigenous or Afrodescendant identity.
- **Marital status** (`LLAVE_SITUACONYUGAL`): Categorical variable distinguishing legal, religious, and informal unions, as well as single, separated, or widowed individuals.
- **Religion** (`LLAVE_RELIGION`): Self-reported religious affiliation, including both institutional and informal beliefs.
- **Primary activity** (`LLAVE_ACTPRIMARIA`): Declared main occupation or economic activity.
- **Commute time to work** (`LLAVE_TIETRASLADO_TRABAJO`): Reported travel time (in minutes) between home and work.

1. See the SIDIE Datasets documentation and methodological notes from Banco de México, available at: `https://www.banxico.org.mx/DataSetsWeb/dataset?ruta=LLM&idioma=en`.

- **Local labor market** (`MERCADO_TRABAJO_LOCAL`): Geographic identifier representing the regional labor market context.

2.2 Data processing

To prepare the dataset for analysis, I excluded variables with more than 10% missing values, while the remaining data underwent imputation using median and mode strategies for numerical and categorical variables, respectively. To simplify interpretation and reduce noise from rare categories, two grouped variables were constructed: `RELIGION_GROUPED`, indicating whether an individual is religious or not, and `ETHNIC_MINORITY`, identifying self-reported Indigenous or Afrodescendant status. The sample was restricted to individuals aged 18 and above with strictly positive labor income, thereby focusing on the economically active population and excluding those unlikely to participate meaningfully in the labor market, such as students or informal workers with undeclared earnings.

To address the strong right-skewness in income distribution and improve model stability, a log transformation (`log1p`) was applied to the income variable as shown in Figure 4. While the dataset is nationally representative, it may under-report informal income. Additionally, excluding individuals with zero income and younger populations could introduce selection bias, and the education variable may not reflect differences in quality or skill acquisition.

2.3 Data Exploration

After preprocessing, the final analytical sample includes 15,996 individuals. To ensure fair evaluation, the data is split into a training set (90%, 14,396) for model estimation and a testing set (10%, 1,600) for assessing out-of-sample predictive performance.

Table 1 summarizes the main characteristics of the training data. Individuals have an average of 8.8 years of schooling and are, on average, 37 years old. Reported income is highly dispersed, with a median of 3,440 pesos and a maximum close to 100,000,000 pesos. This level of skewness introduces potential risks for model training, such as overfitting to extreme values and reduced performance on typical income ranges.

Figure 6 shows the distribution of key demographics. Most individuals are aged 20–50, two-thirds are male, and schooling typically ranges from 6 to 12 years. The majority are married and religious, and about one-third identify as part of an ethnic minority, though many did not specify.

In terms of income patterns, Figure 1 shows that income increases with education

Table 1: Descriptive Statistics of Key Numerical Variables

	ESCOLARIDAD_ACUMULADA (Accumulated schooling years)	**EDAD** (Age)	**INGRESO** (Income))
Count	14,396	14,396	14,396
Mean	8.83	37.19	32,929.40
Std	4.53	13.49	1,666,680.71
Min	0.00	18.00	2.00
25%	6.00	26.00	2,143.00
Median	9.00	35.00	3,440.00
75%	12.00	46.00	6,000.00
Max	23.00	98.00	99,999,999.00

across all years, though the slope flattens or is even negative at higher education levels. Differences in the slopes between years suggest temporal shifts in the returns to education.

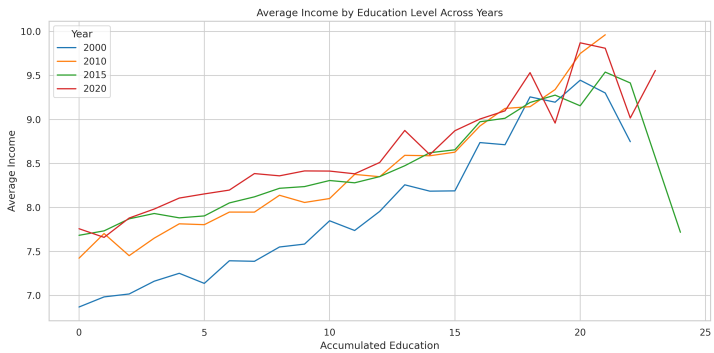

Figure 1: Average log-income by accumulated years of education across census years (2000–2020).

The correlation matrix (Figure 7) highlights moderate associations between accumulated education and income (0.45), while other variables show weak or no linear correlation, suggesting low risk of multicollinearity after the selection of the relevant variables.

Finally, some variables consistently include observations coded as 'Not specified', particularly among categorical variables. These were removed to avoid noisy results. Figure 5 displays the top categories for a subset of variables after grouping, helping to assess the distribution of the remaining unspecified values.

3. Methodology

3.1 Data Preparation and Transformation

To prepare the data for supervised machine learning, I implemented a series of transformations tailored to the structure and quality of the dataset. These steps ensure compatibility with scikit-learn pipelines and enhance the performance and interpretability of the model Pedregosa et al. (2011).

Handling Missing Values: Missing values were addressed using `SimpleImputer`. For numerical variables, including `ESCOLARIDAD_ACUMULADA`, I imputed missing entries using the median of the training set, an approach that is robust to skewed distributions. For categorical variables, missing values were filled with the most frequent category.

Interaction Variable: To capture potential non-constant returns to education over time, I generated interaction terms by applying a polynomial transformation (degree 2) to the combination of `ESCOLARIDAD_ACUMULADA` and `ANIO`, allowing the model to learn whether the returns to education have increased or decreased across different years.

Encoding Categorical Variables: Categorical variables were encoded differently depending on their cardinality:

- **Low-cardinality variables** (fewer than or equal to 10 categories), such as gender and marital status, were encoded using *One-Hot Encoding*.
- **High-cardinality variables**, such as `MERCADO_TRABAJO_LOCAL` (774 unique values), were encoded using *Target Encoding*, which replaces each category with the mean log income for that group.

Scaling Numerical Features: All numerical variables were standardized using `StandardScaler`, which transforms them to have zero mean and unit variance. This step is particularly important for regularized linear models such as Ridge and Lasso.

3.2 Modeling Approach

ML models offer key advantages for income prediction: they flexibly capture non-linear relationships, handle high-dimensional data, and detect complex interactions between variables. Unlike traditional econometric approaches, which prioritize causal inference and

rely on strong parametric assumptions, ML emphasizes predictive accuracy and allows for the exploration of variable importance without imposing rigid model structures Athey (2019). Prior research supports this methodological choice. For instance, (Mareckova and Pohlmeier, 2017) compare traditional statistical techniques with LASSO-based ML variable selection and find that ML outperforms both human and PCA-based selection in predicting labor outcomes. Their findings highlight how ML methods can enhance model performance, especially when dealing with large numbers of potentially relevant predictors.

In this study, a supervised regression framework is used to predict log-transformed income, enabling the assessment of the evolving role of education relative to other demographic and labor market factors. Supervised models are well-suited for this task as they leverage labeled data to learn mappings between features and outcomes. Regression is chosen over classification since income is a continuous variable, and the focus is on modeling variation in levels of income.

To balance predictive performance with interpretability, I implemented five machine learning models: Linear Regression, Lasso and Ridge (regularized linear models suited for feature selection and multicollinearity), and two tree-based ensemble methods—Random Forest and Gradient Boosting—which capture complex nonlinear relationships. All models were evaluated using 5-fold cross-validation. Regularization parameters for Lasso and Ridge were tuned using `LassoCV` and `RidgeCV`, while tree-based models were optimized via grid search on a representative data subset to reduce computation time.

4. Results

4.1 Model Performance Comparison

Table 2 summarizes the performance of the five machine learning models tested, along with a baseline model that predicts the mean income for all observations. Model evaluation was conducted on the test set using three standard metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). MSE captures the average squared difference between predicted and actual values, giving more weight to large errors. MAE represents the average absolute error and is more interpretable and robust to outliers. Finally, R^2 measures the proportion of variance in the dependent variable that is explained by the model, with higher values indicating better explanatory power. These metrics provide a comprehensive view of each model's predictive accuracy on unseen data.

Table 2: Model Performance on Test Data

Model	**MSE**	**MAE**	R^2
Gradient Boosting	0.4804	0.4980	0.4018
Linear Regression	0.4823	0.4984	0.3995
Ridge Regression	0.4823	0.4984	0.3994
Lasso Regression	0.4898	0.5021	0.3901
Random Forest	0.5526	0.5401	0.3119
Baseline (Mean Predictor)	0.8031	0.6726	0.0000

Regarding the relative performance, as shown in Table 2, Gradient Boosting achieved the best overall results with the lowest MSE (0.4804) and MAE (0.4980), and the highest R^2 (0.4018), meaning it explains about 40% of the variance in log-transformed income. Linear and Ridge regression performed similarly ($R^2 \approx 0.3995$), while Lasso was slightly less accurate. Random Forest showed lower predictive accuracy ($R^2 = 0.3119$), and the baseline model, which simply predicted the mean, performed worst with $R^2 \approx 0$.

In terms of absolute predictive error, a MAE of 0.50 in log-income terms implies that predictions deviate from the true values by a factor of about 1.65 on average, that is, predicted income is typically 65% higher or lower than the actual value. For example, a predicted income of 5,000 pesos might correspond to a true value between roughly 3,030 and 8,250 pesos. While this reflects the challenge of modeling income in a highly unequal and segmented labor market, it also demonstrates that machine learning models provide improvements over an average-based approach.

4.2 Feature Importance and Interpretation

To identify which variables contribute most to income prediction, Table 3 presents the top three most impactful features across all models. Based on their performance (see Table 2), I then selected the two best-performing models—Gradient Boosting and Linear Regression—for a deeper analysis. These models were chosen not only for their predictive strength but also for their complementarity: while Linear Regression offers interpretability through model coefficients, Gradient Boosting captures complex nonlinear relationships and interactions with higher predictive accuracy. To evaluate the role of individual predictors, I rely on SHAP (SHapley Additive exPlanations) values. SHAP values offer a unified framework for interpreting feature contributions across both linear and nonlinear models by quantifying the marginal impact of each input feature on a given prediction.

Figure 2 presents the SHAP beeswarm plot for the **Gradient Boosting Regressor**. The x-axis shows the SHAP value, which reflects the contribution of each feature to

the predicted log-income, while the y-axis ranks the features by average absolute SHAP magnitude. The most influential variable is `MERCADO_TRABAJO_LOCAL`, which was target-encoded using average income within each local labor market. High values of this variable (shown in red) are associated with large positive SHAP values, This suggests that the returns to education may have increased over time.

The second most important feature is the interaction term `ANIO` × `ESCOLARIDAD_ACUMULADA`. Here, red-colored points (representing high education in recent years) are concentrated on the right, meaning these values increase the predicted income. This suggests that the returns to education have grown over time: education acquired in more recent years contributes more to income than it did in earlier periods.

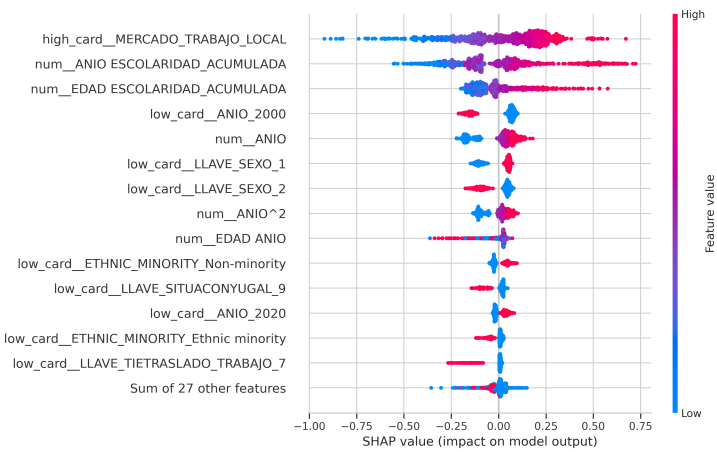

Figure 2: SHAP Beeswarm Plot for Gradient Boosting Regressor

Figure 3 displays the SHAP beeswarm plot for the **Linear Regression**. Interestingly, the results diverge from those of Gradient Boosting. In this case, `MERCADO_TRABAJO_LOCAL` ranks only tenth in importance, while the interaction term `ANIO` × `ESCOLARIDAD_ACUMULADA` and the main effect `ESCOLARIDAD_ACUMULADA` emerge as the top predictors.

The SHAP pattern for the interaction term implies that, contrary to the nonlinear model, education obtained in more recent years may contribute less to predicted income. This is reflected in the clustering of lower SHAP values even for high feature values. Nevertheless, total years of schooling remains a strong positive contributor. These patterns are consistent with the model coefficients reported in Table 5.

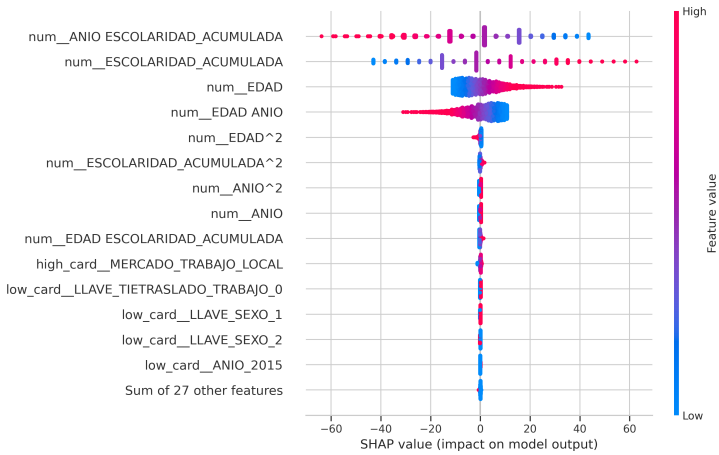

Figure 3: SHAP Beeswarm Plot for Linear Regression

In particular, the divergence in results raises a critical question: **Is income more strongly shaped by structural labor market conditions or by individual demographic characteristics?** Our current analysis cannot fully disentangle these effects. Clarifying this ambiguity would require both a richer set of contextual variables, particularly at the regional and occupational levels, and greater computational capacity. Only then can we better understand the underlying drivers of income inequality in Mexico.

These contrasting results also illustrate the strengths and limitations of different modeling approaches. While Gradient Boosting captures complex interactions and non-linearities, Linear Regression offers clarity and ease of interpretation. This contrast underscores the complexity of the education–income relationship and suggests that the evolving returns to education cannot be fully captured through a single modeling lens. Rather than offering a definitive conclusion, these findings highlight the need for further investigation into how education interacts with broader contextual and temporal factors to shape economic outcomes.

5. Conclusion

Understanding income inequality requires not only access to high-quality data, but also tools capable of uncovering complex patterns within it. This study illustrates how machine learning—combined with interpretability techniques such as SHAP values—can offer a more nuanced view of the drivers of income inequality, particularly in contexts characterized by informality and regional disparities.

The analysis set out to explore whether the importance of education in predicting income in Mexico has increased over time. Using harmonized census data from 2000 to 2020, I applied five supervised regression models—Linear Regression, Lasso, Ridge, Random Forest, and Gradient Boosting—to assess the evolving predictive contribution of education relative to other individual and contextual characteristics.

Results were mixed. Although all models outperformed a naïve baseline, their comparative performance—measured via MAE, MSE, and R^2—showed that Gradient Boosting and Linear Regression performed slightly better, but not substantially so, than Ridge and Lasso. Gradient Boosting identified education and local labor market conditions as key predictors, while Linear Regression revealed a different ranking of feature importance. These discrepancies suggest that the predictive role of education cannot be easily captured by a single modeling framework and may depend on the temporal, geographic, and institutional context.

Overall, the findings indicate that both regional labor market context and evolving returns to education shape income inequality in Mexico. These dimensions interact in complex ways that are not fully captured by traditional methods and may evolve over time in response to macroeconomic and policy changes.

Future research could expand on these results by incorporating household-level variables, more detailed labor market indicators, and richer contextual information—such as sector of employment, firm size, and contractual conditions. Furthermore, leveraging greater computational power and implementing more granular models would allow for a deeper investigation into heterogeneity in income determination, ultimately supporting more effective, evidence-based policy design.

References

Athey, S. (2019). The impact of machine learning on economics. In Agrawal, A., Gans, J., and Goldfarb, A., editors, *The Economics of Artificial Intelligence: An Agenda*, chapter 14, pages 507–547. University of Chicago Press, Chicago.

Becker, G. (1964). *Human Capital*. National Bureau of Economic Research, New York.

Chancel, L., Piketty, T., Saez, E., Zucman, G., Duflo, E., and Banerjee, A. (2022). *World Inequality Report 2022*. Belknap Press of Harvard University Press, Cambridge, MA.

Gomez-Cravioto, D. A., Diaz-Ramos, R. E., Hernandez-Gress, N., and Ceballos, H. G.

(2022). Supervised machine learning predictive analytics for alumni income. *Journal of Big Data*, 9(1):11.

Hausmann, R., Pietrobelli, C., and Santos, M. A. (2021). Place-specific determinants of income gaps: New sub-national evidence from mexico. *World Development*, 146:105566.

Herrera, G. P., Constantino, M., Su, J.-J., and Naranpanawa, A. (2023). The use of icts and income distribution in brazil: A machine learning explanation using shap values. *Telecommunications Policy*, 47(8):102598.

Lopez-Acevedo, G. C. (2006). Mexico: Two decades of the evolution of education and inequality. Research Working Paper WPS3919, The World Bank, Washington, DC. Disclosed in July 2010.

Mareckova, J. and Pohlmeier, W. (2017). Noncognitive skills and labor market outcomes: A machine learning approach. In *Beiträge zur Jahrestagung des Vereins für Socialpolitik 2017: Alternative Geld- und Finanzarchitekturen - Session: Treatment Effects*, number G03-V2, Kiel, Hamburg. ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft.

Matkowski, M. (2021). Prediction of individual income: A machine learning approach. Bachelor's thesis, Bryant University. CC-BY-NC-ND licensed.

Mincer, J. (1974). *Schooling, Experience, and Wages*. National Bureau of Economic Research, New York.

OECD, CIAT, C. and BID (2024). *Estadísticas tributarias en América Latina y el Caribe 2024*. OECD Publishing, París.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Quinn, M. A. and Rubb, S. (2006). Mexico's labor market: The importance of education-occupation matching on wages and productivity in developing countries. *Economics of Education Review*, 25(2):147–156.

A. Additional Material

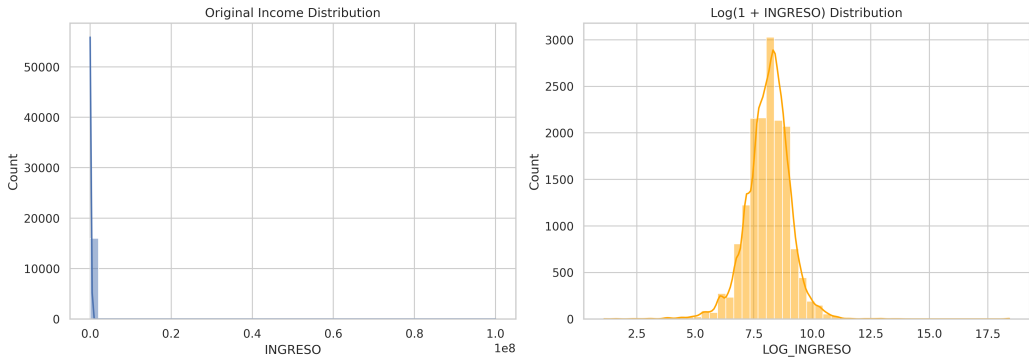

Figure 4: Distribution of log-transformed income

Table 3: Top 3 Most Influential Features for Income Prediction Across Models.

Table 5: Top 10 Most Impactful Features from Linear Regression.

Feature	**Coefficient**	**Abs. Value**
num__ANIO ESCOLARIDAD_ACUMULADA	-21.0687	21.0687
num__ESCOLARIDAD_ACUMULADA	20.8437	20.8437
num__EDAD	7.3695	7.3695
num__EDAD ANIO	-7.0918	7.0918
high_card__MERCADO_TRABAJO_LOCAL	1.1408	1.1408
num__EDAD^2	-0.4068	0.4068
num__ESCOLARIDAD_ACUMULADA^2	0.3420	0.3420
low_card__LLAVE_TIETRASLADO_TRABAJO_7	-0.3166	0.3166
num__ANIO	0.3164	0.3164
num__ANIO^2	0.3164	0.3164

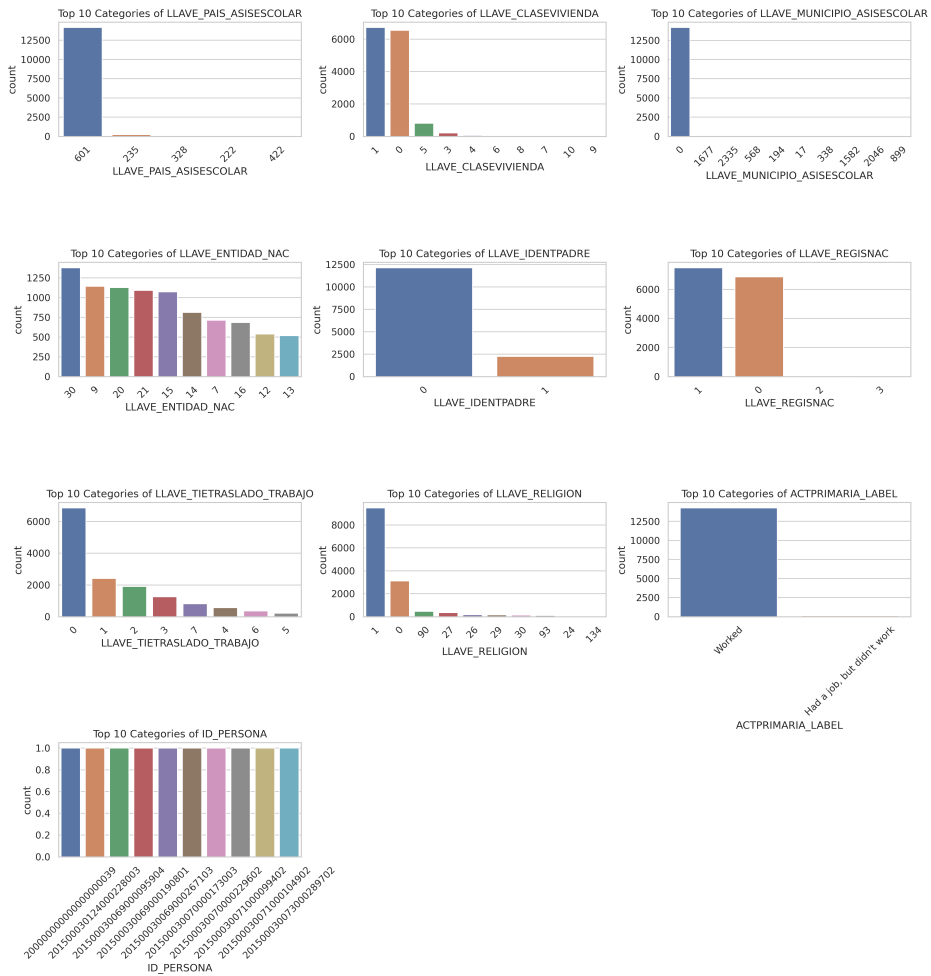

Figure 5: Ten categories across other categorical variables.

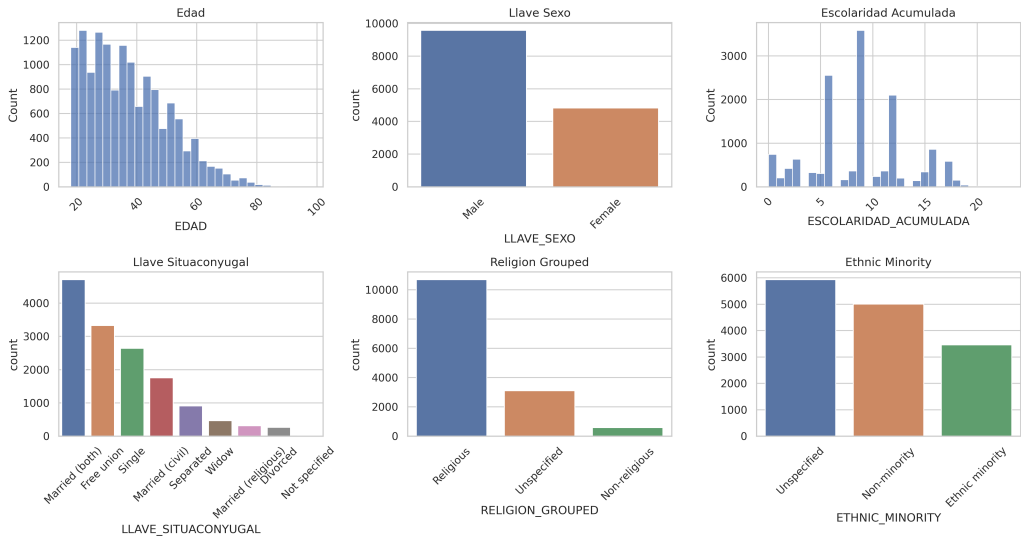

Figure 6: Distributions of key demographic characteristics in the training dataset.

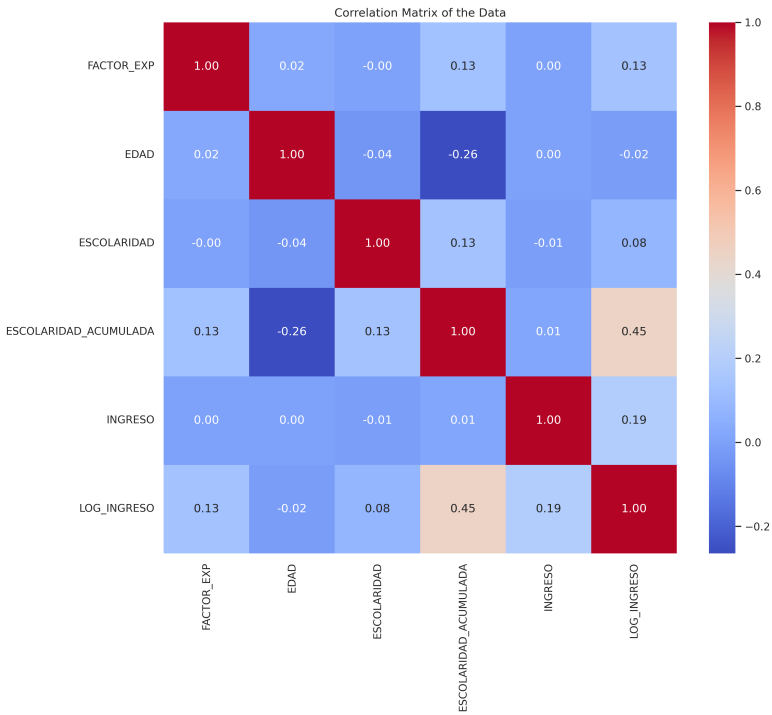

Figure 7: Correlation matrix among selected numerical variables.