Income forecast for conditional cash transfers in Argentina

Cerra, Mariana Montserrat
Cruz, Juan Esteban
Lorcher, Léonie

19th January 2022
Aix-Marseille School of Economics

Contents

1 Introduction

The present study aims to forecast the income of Argentinian individuals in the context of the government's strategy to mitigate poverty through conditional cash transfers.

The conditional cash transfer programs in Argentina, such as the Emergency Family Income program (IFE), are designed to provide financial assistance to households and individuals considered to be particularly vulnerable, such as informal workers and those living in poverty. In particular, the IFE program was launched explicitly in response to the COVID-19 pandemic and the resulting economic crisis. The program provides a significant monetary transfer, making it one of the most extensive measures to contain poverty in the context of the pandemic. Overall, these conditional cash transfer programs are an essential aspect of Argentina's social welfare system, in the context of 36.5% Argentinian individuals living below the poverty line in the first semester of 2022[1]. They demonstrate the government's efforts to address economic inequality and poverty in the country and provide an opportunity to study better ways of targeting vulnerable people in such programs.

The informal labor market, which includes those who work in unregistered or unregulated jobs, often lacks the formal records and documentation required to determine the income of households. This makes it challenging to identify families most in need of support and accurately assess their eligibility for public programs. As economists, we recognize that targeting public policies is highly important for their effectiveness and efficiency in allocating scarce resources. Using algorithms to target individuals or households for public programs is an area of growing interest, as it allows for identifying individuals or families most in need.

In the case of Argentina, where the informal labor market represents a significant share of the workforce (37,8% between April and June 2022[2]), the use of machine learning techniques can improve the targeting of public policies by providing a more accurate prediction of income and other relevant variables. This can help to reduce the trade-off between exclusion error, or the failure to include eligible individuals in the program, and inclusion error, or the inclusion of ineligible individuals. Ultimately, the significance of this study lies in the potential to contribute to the ongoing effort to efficiently allocate resources and mitigate poverty in the context of an informal labor market in Argentina. The methods and findings of this study can be applied in other countries facing similar challenges and serve as a starting point for further research in this area.

In the literature, several studies have investigated machine learning techniques to improve the targeting of public policies, particularly in the context of conditional cash transfers.

For instance, research by Chen (2018) discovered that machine learning algorithms such as CART, C4.5, LASSO, Random Forest, and Adaboost were able to develop prediction models for a cash transfer trial run by the Progresa program in Mexico. His findings revealed that Random Forest and Adaboost possessed the highest level of accuracy for individual outcomes. Furthermore, machine learning models were more effective than traditional structural econometric models when dealing with large amounts of data.

Matkowski (2021) employed data from the U.S. Bureau of Labor Statistics' Population Survey to exhibit the relative superiority of machine learning models in economics through income prediction. The study showed that machine learning methodologies outperformed traditional OLS models with variable selection in prediction. Additionally, the study revealed that income could be predicted with a moderate accuracy level through these models. The best models, gradient boosting and random forest, accounted for nearly 70% of the variation in income with the unique characteristic features available.

Caballero (2021) employed an XGBoost model to predict poverty and unemployment conditions in individuals in Colombia. The model's performance was satisfactory in estimating income, with exclusion and inclusion errors of 44.3% and 27.2%, respectively, which were lower than those presented by Colombian institutions, which were estimated to be 51.8% and 58%. The author also used the Shap technique to explain the correlation between the characteristics used in the predictive models

[1]Source: Instituto Nacional de Estadística y Censos de la República Argentina (INDEC)
[2]Source: INDEC

and income

Noriega-Campero et al. (2020) demonstrated that a shift towards using AI methods in poverty-based targeting could substantially increase accuracy. In particular, for Colombia, the gradient boosting method reduces its exclusion and inclusion errors to around 30%. And for Costa Rica, it reduces both its exclusion and inclusion errors to approximately 26%. This improvement enables an increase in coverage of the poor of nearly a million people in the two countries studied without increasing their social budgets.

Bonaglia (2019) analyzed the targeting mechanisms used by the conditional cash transfer programs in Uruguay and Costa Rica. The author found that Random Forest and Stochastic Gradient Boosting had significant advantages in reducing exclusion and inclusion errors compared to econometric methods used by the Ministry of Social Development to define eligibility. In the case of Random Forest, which had better performance, an 11.3% reduction in errors was achieved compared to the Probit model.

Overall, these studies suggest that machine learning techniques can be an effective tool for improving the targeting of conditional cash transfer programs and identifying households that are most in need of assistance. However, while there is a growing body of literature on the use of machine learning techniques to improve the targeting of public policies, these studies focus on countries with different characteristics than Argentina, such as a smaller share of the informal labor market. This represents a gap in the literature, as the aspects of the informal labor market can significantly affect the accuracy of the targeting. Therefore, our study aims to fill this gap by analyzing the case of Argentina, which poses a unique challenge for obtaining accurate income data. Additionally, there is also a need for interpretability in the literature, as it is crucial for policymakers to understand the most relevant factors in determining eligibility for a program. Our study addresses this need by using SHAP values, which provide interpretability and assess the impact of each variable on the prediction to ensure a better understanding of the model and its results.

To address the question of forecasting the income of Argentinian households, we employ an empirical analysis using machine learning techniques. To conduct our research, we use a dataset from the Permanent Household Survey in Argentina for the 2nd trimester of 2022, led by the National Institute of Statistics and Censuses of the Argentine Republic. This dataset originally contained 50.614 observations and 243 individual and household features. Our approach includes data collection and preprocessing, dimensionality reduction using Lasso and Elastic-net, application of machine learning models such as Support Vector Machine and Random Forest, and model evaluation using prediction accuracy. Our goal is to compare these models to assess the trade-off between exclusion and inclusion errors, which is crucial for policymakers in determining the most efficient allocation of resources. We also use SHAP values to provide interpretability and assess the impact of each variable on the prediction. By employing this empirical analysis, we can develop models that can accurately predict the income of households in the informal labor market, which is critical for efficiently allocating resources to those most in need. This approach allows us to provide a comprehensive and robust analysis of the problem at hand and to draw meaningful conclusions and recommendations for policymakers.

In our study, we found several key results that provide valuable insights for policymakers and practitioners in the field of conditional cash transfers. Our results showed that using machine learning techniques significantly improved the targeting of conditional cash transfers in Argentina. The Support Vector Machine model performed best in terms of prediction accuracy and the trade-off between exclusion and inclusion errors. Additionally, SHAP values provided interpretability and showed that variables related to access to health coverage, adequate housing, employment opportunities, and education were among the most critical factors for income prediction. Our study is unique in its focus on the specific case of Argentina, and our findings have the potential to provide valuable insights into the field of conditional cash transfers, as well as for other public policies that aim to target individuals or households based on their socio-economic characteristics. Ultimately, our goal is to contribute to the ongoing effort to allocate resources efficiently and mitigate poverty in the context of an informal labor market in Argentina.

This study proceeds as follows. Section 2 introduces a detailed overview of the dataset used in the analysis and describes the techniques and algorithms employed to analyze the data. Section 3 presents the results—precisely, an exploration of the data and the results of the predictive models. Finally, section 4 summarizes the study's main findings and discusses their implications.

2 Materials and Methods

2.1 Data presentation

In the process of obtaining and preparing the data, several steps were undertaken to ensure the accuracy and reliability of the results. The data source is the Permanent Household Survey (EPH), which is collected quarterly by the National Institute of Statistics and Censuses of the Argentine Republic (INDEC). This survey constitutes the primary source of information to monitor the socio-economic conditions of the Argentine population, as it covers information on household composition, housing, education, employment, migration, and income, among other dimensions. However, it is essential to note that the EPH is subject to certain limitations, specifically in its coverage of only urban areas with populations exceeding 100,000 inhabitants. As a result, the findings and conclusions of this study are restricted to the areas covered by the survey, and it should be acknowledged that the results may not accurately reflect the conditions in rural areas or smaller towns and settlements.

The initial dataset was composed of two databases, one for individuals with 50,614 observations and 177 features, and one for households with 17,367 observations and 88 features. A merge process was executed, in which the household indicators were paired with those of the people who live there, resulting in the study's unit of analysis being the individual. Exploiting differences between age groups and sex was an advantage of working at the level of individuals.

Subsequently, we created the target variable "income_index" using the programming language R and the "eph" package, adhering to the methodology of INDEC. R was chosen as the programming language for this task due to the availability of the package "eph", which includes tools for downloading and manipulating the Permanent Household Survey data from Argentina. The use of the methodology of INDEC, ensures that creating the "income_index" is executed with precision and rigor, adhering to established statistical methods and providing valid and reliable results. The income index is obtained by dividing the total income of a household (TIH) by the Total Food Basket (TFB)[3] for each household. The detailed steps for calculating it are as follows:

1. We first must consider that the EPH micro database does not publish the month corresponding to each individual record. Therefore, our estimates must be based on quarterly baskets. To achieve this, we calculate a quarterly value of the basket for a unit of equivalent adult (UEA) belonging to each of the six regions in the country using the monthly data of the TFB published by INDEC.

2. Next, we calculate the corresponding unit of equivalent adult for each individual using the equivalence table provided by INDEC. It should be noted that INDEC identifies the reference consumer unit, or equivalent adult, as the age group with the highest concentration of the active population, typically men between the ages of 30 and 60.

3. We then sum the number of equivalent adults for each household to have weighted results. The sum of UEA is then multiplied by the value of the TFB corresponding to one equivalent adult, to estimate the total value of the TFB for each specific household.

4. Finally, for each household, the value obtained from the respective TFB is compared with the total income of that household. If the family has enough income to cover the value of the TFB, then it is not considered poor. If their income is less than the value of the TFB, then that household and the individuals within it are considered to be in poverty. To compute the income index, we divide TIH

[3]The total basic basket (TBB) expands the basic food basket (BFB) by considering non-food goods and services such as clothing, transportation, education, health, housing, etcetera. The basic food basket (BFB) is the set of foods and beverages that satisfy nutritional, kilocalorie and protein requirements, whose composition reflects the consumption habits of a reference population, that is, a group of households that covers these consumption habits. For statistical purposes, the TBB is used in Argentina as a reference in establishing the poverty line.

by TFB, where TFB is specific for each household. Therefore, if the income_index is greater than 1, an individual is considered to not be in poverty, while if the income_index is less than 1, it is deemed to be in poverty.

For example, consider a household with four members from the north region, made up of a 35-year-old man, a 31-year-old woman, a 6-year-old son, and an 8-year-old daughter, equivalent to about 3.09 consumer units or equivalent adults. For January 2020, the value of the TFB in the north region was $13,065 (in Argentinian pesos) per equivalent adult. Then, for this specific household, the TFB amounts to $40,370 ($13,065 x 3.09 adult equivalents). This means that this household needed a total income higher than $40,370 to cover the value of said TFB and not be considered poor. (INDEC)

Once all the steps mentioned above had been completed, the data was moved from R studio to Python to continue with the rest of the project. We proceeded then with the data cleaning. In this stage, we dropped off some variables that were not interesting and renamed the remaining ones. It was decided to carry out an a priori selection of variables due to how inefficient -for descriptive analytics, dimensionality reduction, and machine learning models- it could be to include all of them. In addition, only some variables had relevant information for the study. In summary, we decided to group the selected variables into four categories of variables:

- People's characteristics: includes variables that describe the person such as their age, sex, marital status, and where they live.
- Housing characteristics: comprises variables that describe the type of dwelling in which they live, such as if they have a roof, a bathroom, and a sewage system.
- Socioeconomic characteristics: covers variables that describe their situation in the labor market and their social security and property ownership where they live.
- Sources of income: focuses on the main source of people's income in the last few months.

A detailed description of the variables selected manually can be seen in the appendix at the end of this study.

We also faced the problem of non-response to income, a common phenomenon in household surveys. This phenomenon is because, in the EPH survey, individuals and households are selected randomly and not through self-selection (i.e., individuals do not volunteer to be surveyed). Therefore only some of those selected will be willing to provide complete responses to all questions. It has been observed that affluent households may be more reticent to participate in surveys than other households. Possible explanations for this phenomenon include a reluctance to reveal sensitive information on income and wealth to interviewers or a greater difficulty contacting these households (ECB Statistics Paper No 15, 2016). However, this is a minor issue for the present study, as our primary focus is on correctly identifying impoverished individuals. Since non-response of income is more likely to occur among high-income individuals, the decision was made to remove these observations from the dataset. This resulted in the exclusion of 21.8% of the observations.

In addition to these procedures, we also restricted the age variable to individuals 18 years of age or older. Furthermore, we calculated the age-squared to account for potential non-linearities between age and the income index. It is common for the relationship between age and our feature variable to differ for individuals of different ages. Additionally, we created a new variable, the "overcrowding_index" [4], which was calculated by dividing the number of individuals in the household by the number of rooms designated for sleeping in the dwelling. We further cleaned the remaining variables by eliminating any codes associated with non-responses.

We conducted a thorough examination to detect outliers within the numerical variables by creating box plots for each variable. Based on the results of these plots, we proceeded to remove any extreme values or data that did not conform to common sense. Lastly, we carried out a One-hot encoding by creating dummy variables to select variables and ensured that these dummy variables did not contain

[4]This variable is included in the Appendix tables as a housing characteristic

any outliers and only took on two unique values, 0 and 1. The final cleaned dataset consisted of 26,993 observations and 160 features, ready for further analysis and modeling.

Additionally, it is essential to mention that for the following steps that correspond to the reduction of dimensionality and the estimated models, the logarithm of the "income_index" was used as the target variable. Applying the logarithm would allow it to be normalized, given the presence of skewness to the right. Additionally, one was added to maintain the poverty threshold for those data points that were below one:

$$y = log(income\_index) + 1$$

2.2 Description of the methods

For this work, five methods were mainly used: dimensionality reduction by Lasso, dimensionality reduction by Elastic-net, Random Forest regressor, Support Vector Machine regressor, and Shap values.

2.2.1 Dimensionality reduction by Lasso

It is crucial to note that the database utilized in this study comprises a substantial number of potential regressors. A high number of variables can entail a significant computational burden and may introduce noise by including variables in the model that do not significantly contribute. To address this, shrinkage methods are employed, given that they allow for fitting models with all predictors while incorporating constraints or regularizations that shrink certain estimated coefficients toward zero. James et al. (2021) state that the most well-known methods in this regard are Ridge and Lasso. These methods aim to reduce the variance and enhance the model's fit.

In contrast to Ridge, Lasso's method seeks to force some estimated coefficients to be precisely zero when the penalty parameter is sufficiently adequate. In contrast, Ridge's method sinks certain coefficients toward zero without reaching an exact zero value. Consequently, Lasso performs a variable selection in which it obtains coefficients β_λ^L that can be obtained from the following minimization (James et al., 2021):

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

The minimization involves the selection of a penalty parameter "λ". The choice of "λ" is of paramount importance, as a very low parameter will penalize only a minimal number of coefficients. At the same time, a very high value will yield a null model in which all coefficients are zero. According to James et al. (2021), this parameter choice must be made by the smallest cross-validation error. First, by choosing a grid of possible values for the parameter, and then computing the cross-validation error of each value.

During the execution of this selection of variables, it was not only important to use the models, but also GridSearchCV was used. GridSearchCV seeks to select the best and most optimal parameters based on a grid that considers different possible values for the Lasso hyperparameters. The selection of the parameters works with cross-validation, a method that seeks to divide the data into subgroups (k-folds), k-1 are used to train the model, and 1 is used to test it. In cross-validation the process is repeated several times testing all the combinations of test/train sets and the average error is used as the indicator of optimization since it is desired that this is the smallest possible. The hyperparameters that from cross-validation obtain the smallest errors will be selected, which shows that they are the best values to model the data.

2.2.2 Dimensionality reduction by Elastic-net

The Elastic Net dimensionality reduction method is a combination of both Lasso and Ridge methods. Specifically, the Elastic Net modifies the minimization, incorporating a penalty term in the following manner:

$$RSS + \lambda(\sum_{j=1}^{p}(\alpha\beta_j^2 + (1-\alpha)|\beta_j|))$$

Hastie (2005) stated that the Elastic Net aims to select variables like Lasso and shrinks the coefficients of correlated variables like the Ridge method. Elastic Net will generally result in more non-zero coefficients compared to Lasso reduction. However, some of these coefficients will have a relatively small magnitude. Similar to the Lasso method, the Elastic Net also involves selecting the penalty parameter "λ", but additionally requires the selection of the hyperparameter "α". Specifically, "α" defines the degree to which the penalty term is a mixture between Ridge and Lasso. In simpler terms, with the Elastic Net method, a reduction similar to Lasso can be achieved by setting "α" equal to 0, and a reduction similar to Ridge can be achieved by setting "α" equal to 1.

For elastic-net, GridSearchCV is also used as a hyperparameter optimization method. The difference with respect to Lasso is that while Lasso is only looking for one hyperparameter, the elastic-net method seeks to optimize two hyperparameters as just shown in the explanation of the process.

2.2.3 Random Forest regressor

To comprehend a Random Forest's intricacies, one must first familiarize oneself with tree-based methods. These techniques are useful for both classification and regression problems. For the purposes of this analysis, we shall focus on using regression trees with the ultimate goal of predicting the income index.

A decision tree is based on the identification of rules to divide the data through a series of division criteria, which creates new nodes with each split. These trees can vary regarding the number of features, the sample size considered at each node, and the depth it can possess. Specifically, for a regression problem, a tree will divide the data into non-overlapping regions, where observations that fall within the same region will yield the same prediction (James et al., 2021).

The construction of these regions is achieved through recursive binary splitting, with the aim of minimizing the residual sum of squares (RSS). The objective is to ensure that the tree is not overly complex, as this can lead to overfitting of the data. To combat this, it is necessary to prune the tree by obtaining a subtree that penalizes complexity (James et al., 2021):

$$\sum_{m=1}^{|T|}\sum_{i:x_i\in R_m}(y_i - \hat{y}_{R_m})^2 + \alpha|T|$$

The selection of the appropriate value for α is achieved through using K-folds cross-validation, with the objective of selecting the tree that minimizes the mean error. Another crucial concept to understand regarding Random Forest is bagging or Bootstrap aggregation. This procedure aims to reduce variance by averaging the results of different predictions obtained from different bootstrap samples.

Given the explanation above of decision trees and bagging, it can be stated that according to James et al. (2021), a Random Forest is an ensemble method that combines several simple "building block" models to create a more powerful model. The Random Forest utilizes bagging and, additionally, a random split of predictors. A sample of predictors is used for each split, preventing the majority of available predictors from being considered. This, in turn, decorrelates the trees to be averaged and significantly reduces the variance of the model.

The execution of the Random Forest model also requires the use of GridSearchCV, which was already explained above, and additionally RandomizedSearchCV was used, which is another method of searching for the optimal hyperparameters for the model. The RandomizedSearchCV method also uses cross-validation on a set of defined values to test each hyperparameter, but as commented on the scikit learn page, unlike GridSearchCV, not all values are tested but a fixed number of parameter settings is sampled.

2.2.4 Support Vector Machine

The Support Vector Machine (SVM) is an extension of the Support Vector Classifier that seeks to expand the feature space to accommodate non-linear boundaries between classes, thereby enabling the identification of the class to which each observation belongs (James et al., 2021). To comprehend the SVM, it is essential to begin with the Maximal Margin Classifier, considered a classifier that separates observations using a hyperplane (a flat affine subspace). The selection of the optimal hyperplane is based on identifying the observations close to the boundary and creating support vectors, then considering the distance to those vectors as the margin. In this vein, the Support Vector Classifier, also referred to as the soft margin classifier by James et al. (2021), seeks to find the smallest possible margin while allowing some observations to be on the wrong side of the margin and even on the incorrect side of the hyperplane.

The SVM utilizes non-linear boundaries by enlarging the feature space through kernels. The concept of using kernels is to quantify the similarities between observations; thus, kernel selection allows for flexibility in the decision boundary. As an example, using kernels for a two-dimensional dataset results in a non-linear function such as:

$$f(x) = \beta_0 + \sum_{i \in S} K(x, x_i)$$

The SVM can be used for more than two class classifications through one-versus-one or one-versus-all classification.

The execution of the SVM model also requires the use of GridSearchCV, a method that tunes the hyperparameters from cross-validation. Additionally, it should be noted that among the methods used to better understand the results of the models, obtaining the confusion matrix is also used. A confusion matrix is a summary of the results obtained after carrying out supervised classification models. Said matrix shows the number of data values that were correctly classified and those that the model classified erroneously. For our analysis, the confusion matrix is very important despite not having a classification problem but one of regression. However, thanks to the fact that our analysis allows us to establish a value as a poverty threshold, this allows us to divide the data between the poor population and the non-poor population. Thus, we can obtain a confusion matrix under the assumption that the prediction of the poverty index can be considered as the criterion that the government would take for the inclusion or exclusion of a money transfer policy that wishes to reduce poverty in the country.

2.2.5 Shap values

SHAP (SHapley Additive exPlanations) values are a unified measure of feature importance that allows for the explanation of any machine learning model. Developed by Lundberg and Lee (2017), SHAP values provide a way to understand the contribution of each feature to the prediction of a specific instance. These values are based on Shapley values from cooperative game theory, which assign a value to each player in a game that represents their contribution to the overall outcome. In machine learning, each feature is considered a player, and the prediction is the overall outcome. The SHAP values indicate the average marginal contribution of each variable, and they guarantee that the sum of all the feature importance values equals the difference between the model's output and the expected value of the output.

SHAP values have several advantages over other feature importance measures. Firstly, they are model-agnostic, meaning they can be used to explain any machine-learning model. Secondly, they are consistent with the model's underlying assumptions, such as linearity or non-linearity. Thirdly, they explain each instance, not just average feature importance across all instances. Finally, they also provide a measure of uncertainty, indicating the degree of feature importance variation across different instances (Lundberg & Lee, 2017).

The SHAP library in Python is an open-source library for computing SHAP values for various machine-learning models (Strumbelj & Kononenko, 2014). It has been widely used in many fields,

such as finance, healthcare, and computer vision. The library can compute SHAP values for a specific instance, a group of instances, or a dataset as a whole. It also provides various visualization tools to help understand the results.

3 Results

3.1 Exploratory analysis

Prior to engaging in data manipulation through dimensionality reduction and subsequent model construction, it is imperative to gain a thorough understanding of the data set to be utilized. To achieve this objective, descriptive statistics were employed. The objective of this methodology is to visually depict the trends of key variables to comprehend the distribution of the data and identify any potential disparities. Additionally, it serves to verify any assumptions made regarding certain variables, as well as to determine the relationship between features and between the target and explanatory variables.

3.1.1 Univariate Analysis

In the first analysis of the descriptive statistics section of the study, we aimed to examine the distribution of the variables themselves, without taking into account relationships between features. The primary focus of this analysis was on the target variable, which is the index. This quantitative and continuous variable is closely related to the categorical variable poor_category. Depending on the value of the index, individuals are classified as indigent, poor, or non-poor.

To begin the analysis, we examined the distribution of the index variable. The below histogram revealed that the distribution was close to a Chi-squared distribution and was heavily skewed to the right. We decided to apply a logarithm transformation (and to add a one to keep the same threshold for poor/non-poor) to the index to address this issue. The resulting distribution of the logarithm of the index was more balanced and had a shape close to a normal distribution. A significant portion of the population had an index between 0 and 2, with 1 being the turning point. Individuals with an index below one were considered poor, while those above one were considered non-poor.

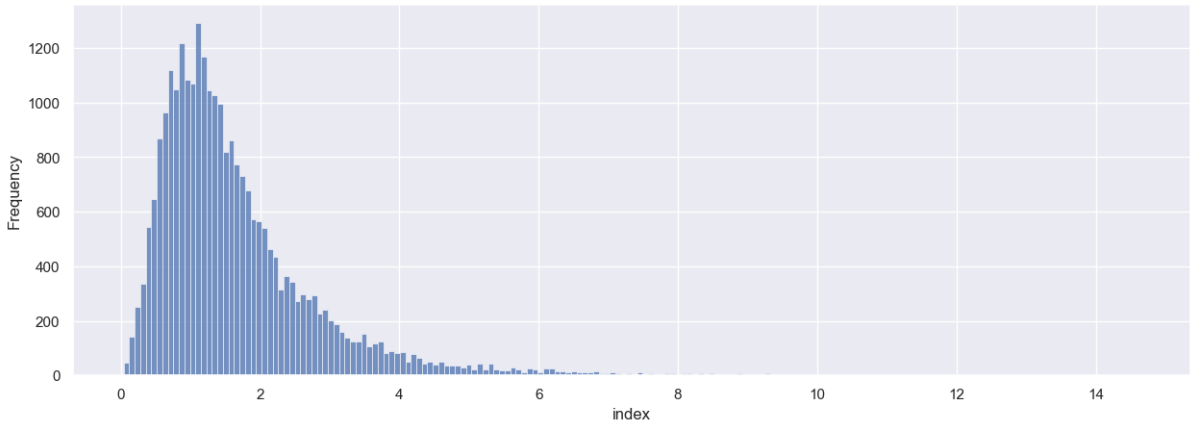

Figure 1: Distribution of the index variable

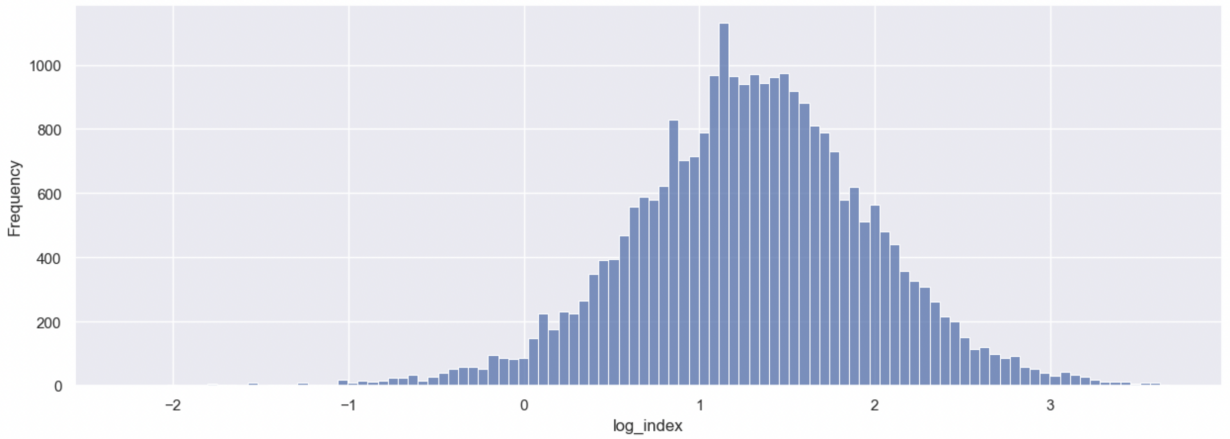

Figure 2: Distribution of the logarithm of the index variable

Below are the summary statistics of these two variables. The logarithm of the index variable, log_index, has a mean value slightly lower than the original but still above 0. Additionally, the standard deviation of log_index is significantly reduced compared to the original index variable. This indicates that the logarithmic transformation effectively reduced the skewness and increased the distribution's normality.

Table 1: Summary statistics for the index and log-index variables

	Index	**Log-index**
count	26993	26993
mean	1.69	1.29
standard deviation	1.30	0.68
minimum	0.03	-2.26
25%	0.87	0.87
50%	1.35	1.30
75%	2.07	1.72
maximum	14.63	3.68

Next, we examined the distribution of poverty classes within the sample in the plot below. The proportion of individuals classified as poor in the sample is relatively high, with 26% of the population falling into this category. Additionally, 5.5% of the sample is classified as indigent, indicating a significant portion of individuals facing severe financial hardship. This indicates that roughly 31% of the sample falls within the poverty class, and it highlights the need for targeted interventions and policies to address poverty within the population.

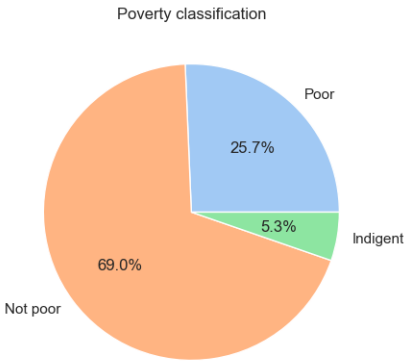

Figure 3: Poverty classification

In terms of the age of the sample, the descriptive statistics and frequency graph indicate that the majority of the sample is composed of active individuals, with 75% of the population being under 59 years of age. The mean age of the population is 45 years old, which is indicative of a relatively young population. The interval with the highest frequency is between 33 and 45 years old, while the interval with the lowest frequency is 75 years and above. This suggests that the population may be relatively young, with a smaller proportion of older individuals present.

Table 2: Summary statistics for the age feature

	Age
count	26993
mean	44.92
standard deviation	18.30
minimum	18
25%	29
50%	43
75%	59
maximum	104

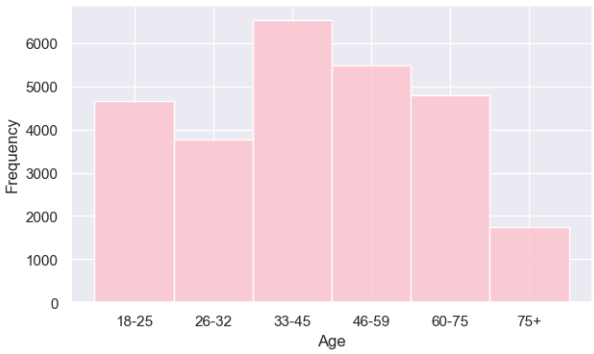

Figure 4: Age frequency by intervals

The sex distribution of the sample reveals that there is a slight majority of females, with a proportion of 53.3% of the individuals identifying as women, while 46.7% identify as men. This distribution is relatively balanced, with only a small difference between the two groups.

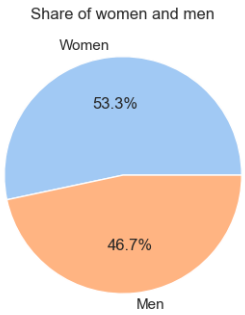

Figure 5: Sex distribution

Concerning the educational level, participants have been asked to choose between the highest level attained:

- 0 (coded 7 in the original dataset) = No instruction
- 1 = Incomplete primary (includes special education)
- 2 = Full primary
- 3 = Incomplete secondary
- 4 = Full secondary
- 5 = Incomplete university degree
- 6 = Complete university or higher

The distribution of educational level among the population in our sample in the two graphs below shows that a quarter of individuals possess a full secondary level of education, with a frequency of 27%. This indicates that a significant portion of the population has completed a basic level of education, which is generally considered necessary for participating in the workforce and pursuing higher education or vocational training. Furthermore, we can observe that the frequencies of individuals with incomplete primary, full primary, and incomplete secondary levels of education are relatively high, with 6%, 16%, and 17%, respectively. This suggests that there may be barriers or limitations in access to primary and secondary education within the population, given that approximately 39% of the sample did not finish secondary school.

On the other hand, we can also observe that the frequencies of individuals with incomplete or complete university degrees are relatively similar, with 16% and 17%, respectively. This may indicate that there are opportunities for higher education within the population, but there may be challenges or obstacles in the transition from secondary to higher education. Additionally, it is worth noting that the frequency of individuals with no instruction is shallow, at only 0.8%. This suggests that efforts or policies may be in place to promote literacy and primary education among the population.

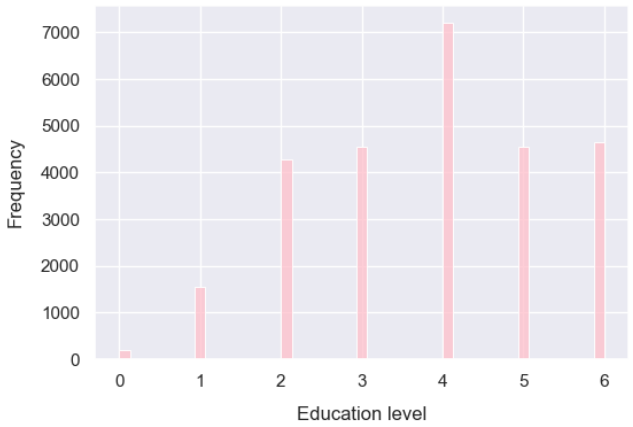

Figure 6: Education level per category

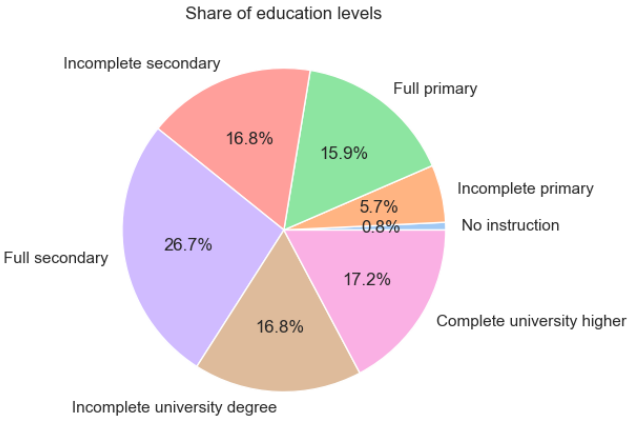

Figure 7: Education level shares

Concerning the size of participating families, we used the feature of "number of people in the household" to get a broader scope in the following graph. The distribution of family size among the sample exhibits a relatively balanced distribution, with the majority of families being composed of two to four individuals. The lowest frequency is observed in families with a size of seven or more individuals, while the highest frequency is found in families with a size of two individuals.

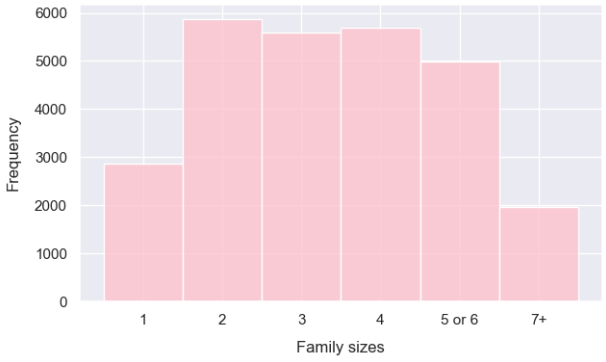

Figure 8: Size of the households

The survey sought to gather information on the sources of income for the individuals in the sample. Participants were asked to indicate whether they had lived from specific sources in the last three months. These sources included work earnings, retirement income, unemployment insurance, social assistance, renting a property, business earnings, investments, and savings. The responses were recorded as either "yes" or "no" for each category. The following is an analysis of the proportion of individuals who reported receiving income from each source.

The majority of the individuals in the sample, 84%, rely on work earnings as their primary source of income. This is not surprising, as it is expected that most individuals would rely on their employment as a means to sustain themselves and their households. However, it is notable that a significant portion of the sample, 42%, also relies on retirement income.

On the other hand, it is also noteworthy that a relatively small proportion of the sample, 0.2%, relies on unemployment insurance as a source of income. This could indicate a lack of access to unemployment benefits. Similarly, only 0.2% of the sample relies on business earnings, indicating that self-employment or entrepreneurship is not a prevalent source of income within the population.

Another interesting insight is that 21% of the sample relies on social assistance as a source of income. This may be indicative of the socio-economic disparities within the population, as social assistance is typically provided to those who are in need of financial support. Additionally, it is also

noteworthy that 2.4% of the sample relies on renting a property as a source of income, indicating a relatively low rate of homeownership within the population.

Lastly, it is worth mentioning that 30.9% of the sample relies on savings as a source of income. This could indicate a lack of other stable sources of income, such as employment or retirement benefits.

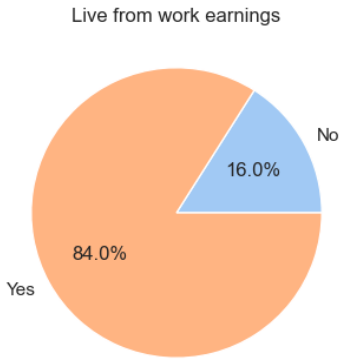

Figure 9: Work earnings

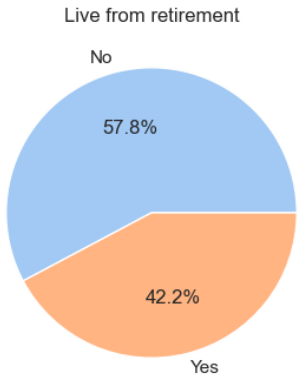

Figure 10: Retirement income

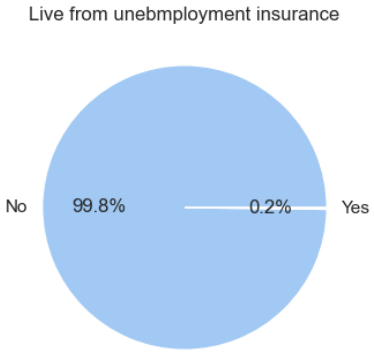

Figure 11: Unemployment insurance

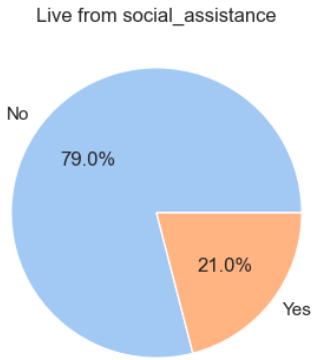

Figure 12: Social assistance

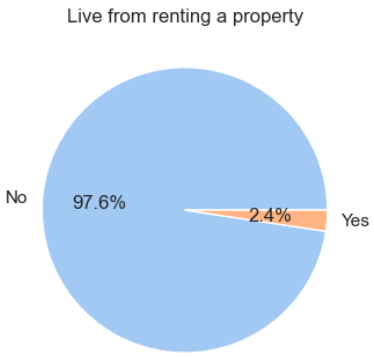

Figure 13: Property renting

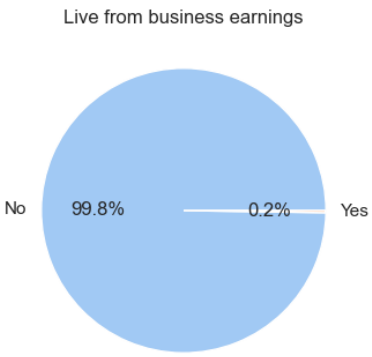

Figure 14: Business earnings

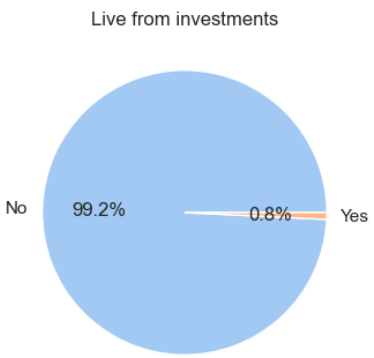

Figure 15: Investments

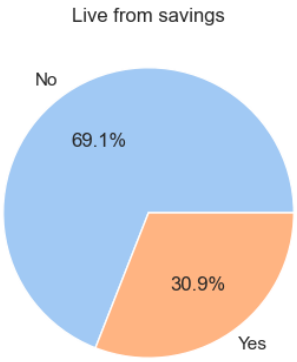

Figure 16: Savings

In terms of regional distribution, it can be observed in the two graphs below that the North-West and Pampa regions are the most heavily represented in the sample, with 26.4% and 26.8% respectively. However, it should be noted that the other four regions - North-East, Cuyo, Patagonia, and Greater Buenos Aires - are also represented in the sample, with roughly similar proportions of around 11-13% each. This suggests a relatively even distribution of the population across these regions.

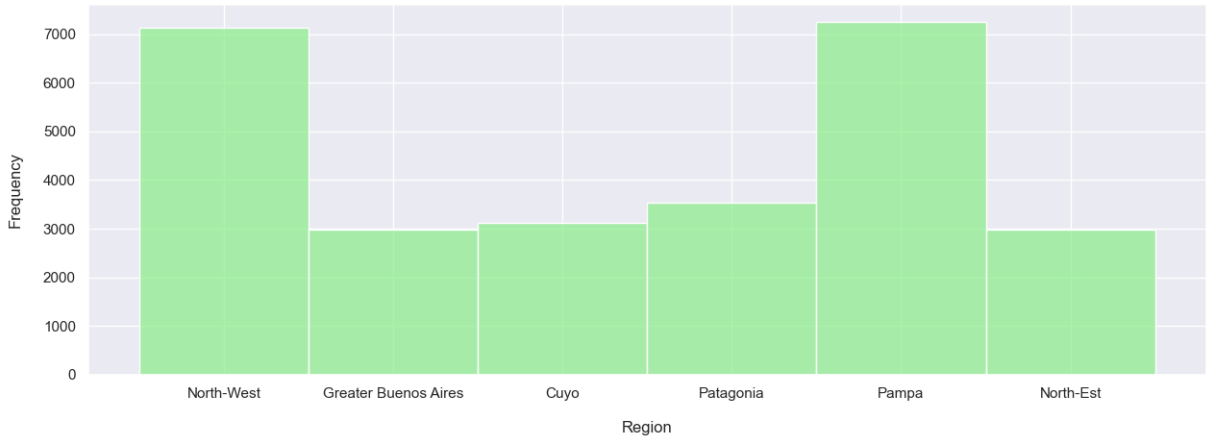

Figure 17: Frequency by regions

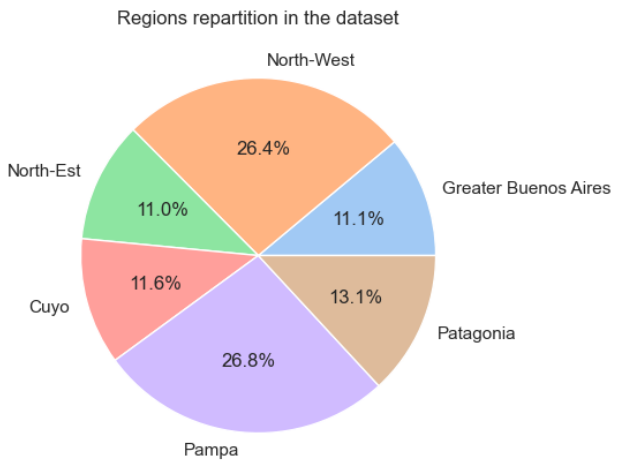

Figure 18: Percentage of individuals in each region

3.1.2 Multivariate analysis

In this second analysis we will examine relationships between variables. First, we aimed to analyse a possible correlation between education and poverty. From the graph below, it is evident that there is a strong correlation between the two variables.

The proportion of individuals with a complete university or higher level of education (level 6) is significantly lower in the poor class, with none of them being categorized as indigent. On the other hand, in the non-poor class, the majority of individuals have at least completed full secondary education (level 4), with a significant proportion having completed university education or higher. This suggests that access to higher education may serve as a protective factor against poverty. This is consistent with the literature on earnings and education, which suggests that higher levels of education are associated with better economic outcomes (e.g. Becker & Chiswick, 1966).

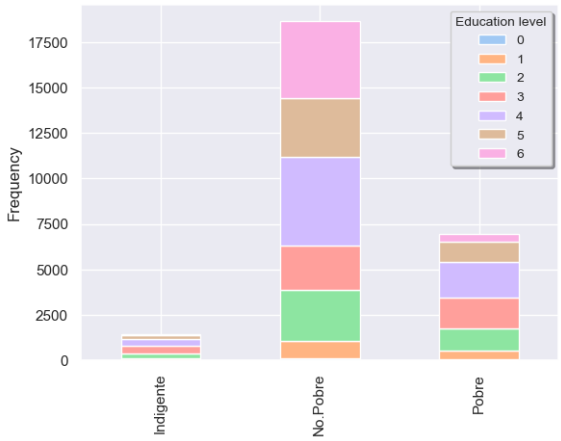

Figure 19: Education and poverty correlation

In the following, we will analyze the relationship between poverty classification and the regions. In figure 20 is evident that the North-West and Pampa regions have a higher frequency of indigent individuals compared to the other regions. However, upon further examination through the use of a second graph created with the software Tableau, it is apparent that the poverty rates in other regions are also relatively high.

It is worth noting that the poverty rate in the Pampa region is relatively close to the rates in the North-West and North-East regions, which is indicative of a relatively high poverty rate in these regions. Additionally, the Patagonia region has a lower poverty rate than the other regions, which could be related to the specific characteristics of this region such as its economic activities and population density. This analysis highlights the importance of considering regional variations in the models we will later develop.

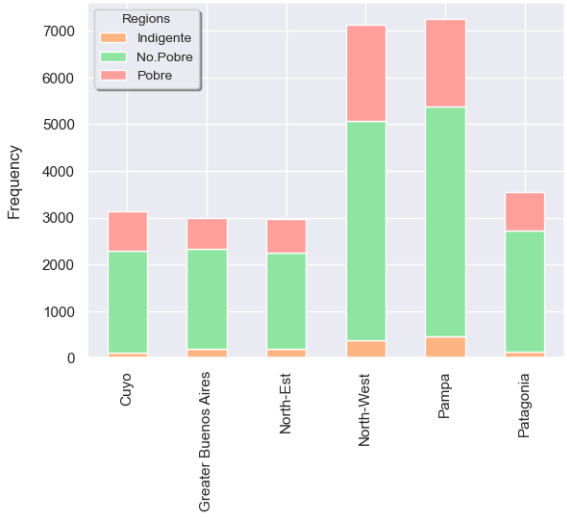

Figure 20: Region and poverty correlation

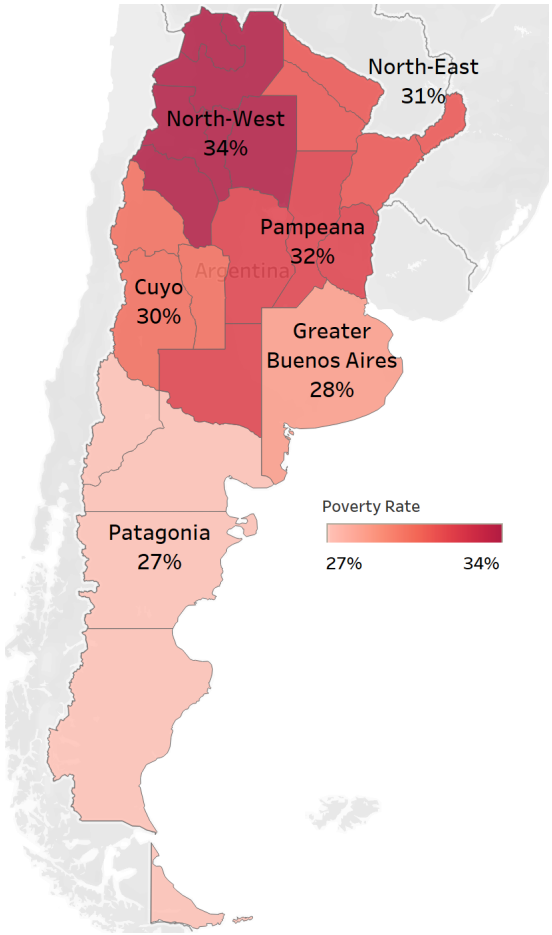

Figure 21: Povery rate by region

3.1.3 Correlation analysis

To study the correlation between variables, we build correlation heatmaps. The correlation matrix provides a quick and easy way to identify the relationships between different quantitative variables. It is important to note that a high correlation coefficient, whether positive or negative, does

not necessarily imply causation, but rather a correlation or association between the variables. Each coefficient of the following table corresponds to the degree of correlation. A correlation coefficient of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable also increases. A correlation coefficient of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases. A correlation coefficient of 0 indicates that they are independent from each other. The colors might help to identify the correlated variables faster.

First, we present the correlation matrix for the quantitative variables:

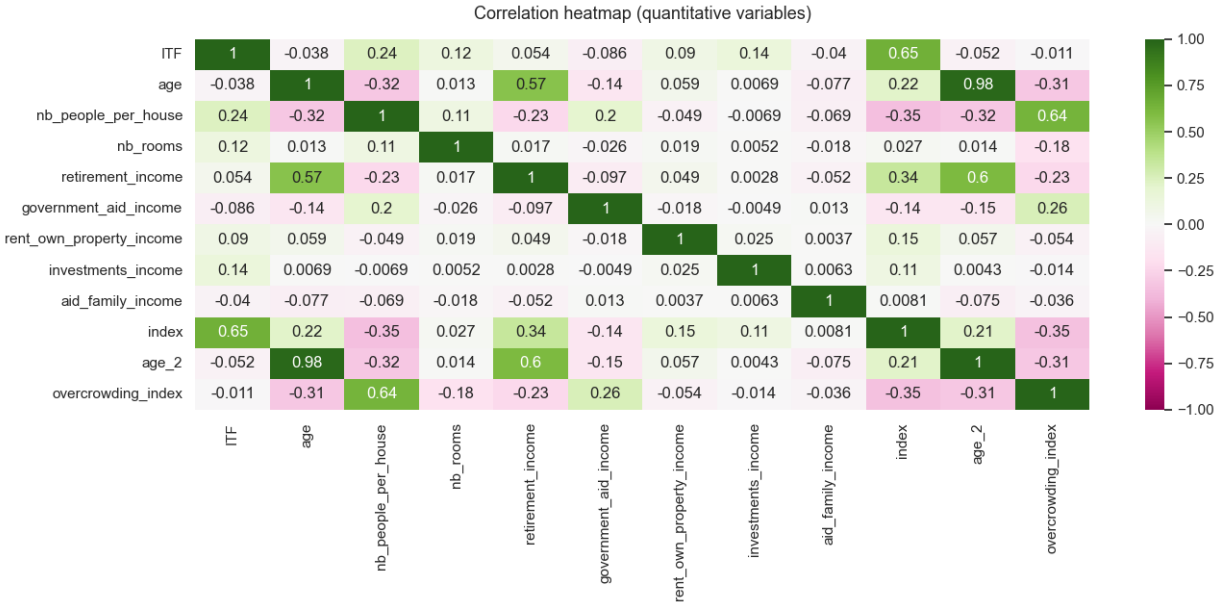

Figure 22: Correlation Heatmap for quantitative variables

From it, we can see that the variable 'index':

- is positively correlated with the variable 'age' and 'retirement income', which is expected as older individuals are more likely to have retired and thus have a higher retirement income, which in turn would likely result in a higher index.
- is negatively correlated with the variable 'number of people in the house' which could indicate that households with higher income tend to be smaller.
- is also negatively correlated with the amount of government aid received and with the overcrowding of housing, which could indicate that those who have a higher income are less likely to rely on government aid and have less crowded housing.
- is positively correlated with the amount earned from the rental of a property and from investments. This could suggest that those with higher income are more likely to have additional sources of income from rental properties and investments.

We then built the correlation heatmap for binary variables, which provides a visual representation of the relationship between the different binary variables in the dataset. From the heatmap, we can see that the categorization as "poor" is positively correlated with the probability of living off social assistance. This suggests that individuals who are classified as "poor" are more likely to rely on social assistance as a source of income. Similarly, the categorization as "indigent" is also positively correlated with the probability of living off social assistance. While the category "non-poor" is negatively correlated with this feature. This highlights the important role that social assistance plays in addressing poverty and indigence in the population.

Figure 23: Correlation Heatmap for binary variables

3.2 Predictive analysis results

3.2.1 Dimensionality reduction

In this stage of the study, we aimed to perform a dimensionality reduction on the cleaned dataset to improve the performance and interpretability of the models that will be applied in the following steps. The first step we took was calculating the correlation between the variables in the dataset. We found a percentage of 6.25% of the variables had a correlation higher than 80%. A high correlation between variables can lead to multicollinearity, which can negatively impact the performance and interpretability of the models. To address this issue, we proposed using two dimensionality reduction techniques: Lasso and Elastic-net.

Lasso is a linear model that adds a regularization term to the cost function, which is a sum of the absolute values of the coefficients. This regularization term is controlled by a hyperparameter, alpha, which determines the strength of the regularization. The advantage of Lasso is that it can shrink some coefficients to zero, effectively removing them from the model. This model can be helpful in situations where there are a large number of variables, but only a subset of them are relevant to the model. However, it can also lead to underfitting if the regularization is too strong, as it may remove essential variables from the model.

Elastic-net is a combination of Lasso and Ridge regularization. Like Lasso, it shrinks some coefficients to zero, but it also shrinks all coefficients by the same amount, preventing overfitting. Elastic-net is controlled by two hyperparameters: alpha and l1_ratio. Alpha controls the strength of the regularization, and l1_ratio controls the proportion of L1 regularization in the combination. This combination of L1 and L2 regularization can be useful when the data contains a mix of correlated and uncorrelated variables.

Before running Lasso and Elastic-net, we scaled the data. Scaling the data is essential because these methods are sensitive to the scale of the variables. By scaling the data, we ensure that all variables are on the same scale, allowing regularization to be applied uniformly across all variables and preventing some variables from dominating the model.

We used the GridSearchCV function from the scikit-learn library to optimize the hyperparameters of the Lasso and Elastic-net models. GridSearchCV is a powerful method that performs an exhaustive search over a specified parameter grid, and it selects the best set of hyperparameters based on cross-validation performance. We optimized the hyperparameters 'alpha' for Lasso, and 'alpha' and 'l1_ratio'

for Elastic-net. We fitted the GridSearchCV objects to the data, and the best hyperparameters selected by the grid search are shown in Table 3.

Table 3: Best hyperparameters selected by the grid search

Hyperparameter	**Lasso**	**Elastic-Net**
alpha	0.001	0.001
L1_ratio		0.54

We checked the percentage of coefficients shrunk to zero; 17% in the case of Lasso and 13% for Elastic-net. We then compared the Lasso and Elastic-Net models using cross-validation and R-squared. The results showed that Elastic-net was slightly better, with a mean cross-validation score of 0.5060 and an R-squared of 0.5180. We proceeded in the following sections with Elastic-net due to its considerable advantages over Lasso and its ability to handle correlated variables effectively.

As a result of the Elastic-net approach, 20 variables were shrunk to zero, leaving a total of 139 variables in the dataset. These variables correspond mainly to people's characteristics and housing characteristics, such as floor type and access to water.

Regarding people's characteristics, the variables that were removed include education level of high school completed, marital status of widow/widower, the household role of the spouse/partner of the boss of the household, being a father-in-law, living in the regions of Gran Buenos Aires, Cuyo or Patagonia, and living in the provinces of Río Gallegos or Río Cuarto.

Concerning housing characteristics, the variables that were taken out include floor type of mosaic/tile/wood/ceramic/carpet, housing type of tenancy room, fuel for cooking gas, a bathroom with exclusive use of the home or shared with another household from the same dwelling, bathroom drain to the public network or the septic chamber, and access to water outside the house. These variables may not be as important as other factors in predicting the model's outcome, or they may be highly correlated with other variables in the dataset. It is worth noting that variables inside the categories of socioeconomic characteristics and related to income and living sources were not removed by the Elastic-net approach. This result suggests that these variables are essential for the model and are not highly correlated with other variables.

With the dimensionality reduction applied, this dataset will be used for performing the Random Forest and the Support Vector Machine models. By removing redundant and less important variables, the models will be able to learn more effectively and will be easier to interpret.

3.2.2 Random Forest

For the Random Forest model, it was initially defined how the set of data would be divided. It was decided that 80% of the data was the training set, 10% the validation set, and 10% the test set. It was decided to divide the data into three sets since the training set, which is the largest, will allow the model to be run with the selected hyperparameters. The validation set will allow us to adjust the hyperparameters with which the model was run through the results on the Mean Squared Error (MSE), and if these results are allowing to obtain the best possible results (which translates into a low MSE compared to the train set). Finally, the test set will enable evaluation of the performance of the selected model since it seeks to compare how the prediction of our model would be with respect to what should actually be obtained.

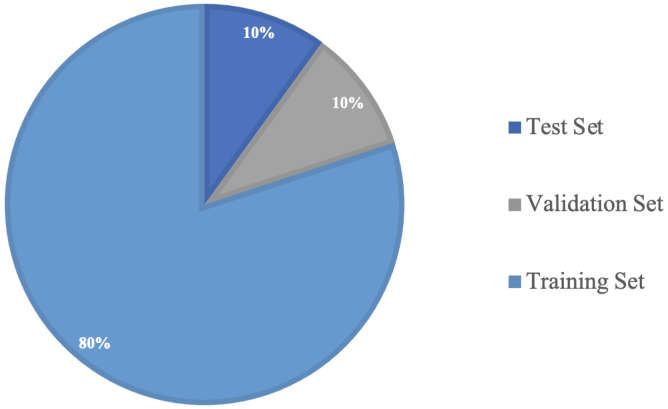

Figure 24: Train / Test / Validation split

Before carrying out the models, it is essential to mention that the dataset was rescaled by standardization for all the explanatory variables since this is a necessary process for specific models to be compared. A first Random Forest Model model is estimated without tuning the hyperparameters but using the predefined hyperparameters of the 'RandomForestRegressor' package of scikit learn in Python. The hyperparameters that are sought to be adjusted in this model are the following:

- n_estimators: refers to the number of trees to be estimated in the Random Forest. By default is 100.
- max_features: refers to the maximum number of features considered at every split. By default is the total number of features that the dataset possesses.
- max_depth: refers to the maximum number of the levels in each tree. By default, the nodes are expanded until all leaves are pure or until all leaves contain less than the minimum number of observations required to split a node.
- min_samples_split: refers to the minimum number of observations required to split a node. By default is 2.
- min_samples_leaf: refers to the minimum number of observations required at each leaf node (node without successors). By default is 1.
- bootstrap: refers to the method of selecting samples for training each tree. By default, bootstrap samples are used when building trees.

The predefined set of hyperparameters can generate a good model with the training sample. However, many default hyperparameters are developing a model overfitting the data. The model can be overfitting as is using a min_samples_split and min_samples_leaf very small, generating a particular model that fits the training sample perfectly but does not perform well on the validation and test sample. This first model returns an MSE of 0.027 for the training sample and 0.172 for the validation sample. The MSE is higher for the validation sample compared to the training sample. This shows that the model needs to use better hyperparameters as it can be overfitting; therefore the hyperparameters need to be tuned.

For tuning the hyperparameters, two approaches are used. Firstly, we use the RandomizedSearchCV. This method allows searching for the best hyperparameters through a set of values established for each parameter (a grid). In our case, the following values were taken to be considered for the grid search:

The objective of this analysis is to optimize the performance of the Random Forest model through the selection of appropriate hyperparameters. To achieve this, we employed a grid-search strategy

Table 4: RandomizedSearchCV grid

Hyperparameter	**Values**
n_estimators	200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000
max_features	'auto', 'sqrt'
max_depth	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None
min_samples_split	15, 25, 45
min_samples_leaf	10, 20, 40
bootstrap	True, False

utilizing 3-fold cross-validation over a set of ten predefined parameter settings. While increasing the number of iterations and cross-validation folds may yield more comprehensive results and reduce the risk of overfitting, it is important to consider the trade-off with computational time. The results of the initial grid-search are presented in the table of selected hyperparameters. To further validate the results, a second iteration of the grid-search strategy was implemented using 3-fold cross-validation, this time focusing on the parameter values surrounding the optimal values obtained from the initial search. The final results of this second grid-search are also displayed in the following table:

Table 5: Selected hyparameters of each grid-search

Hyperparameter	**RandomizedSearchCV**	**First GridSearchCV**	**Second GridSearchCV**
n_estimators	800	700	600
max_features	124	100	90
max_depth	100	100	100
min_samples_split	15	10	10
min_samples_leaf	20	15	10
bootstrap	True	True	True

We are achieving stable outcomes for 'min_samples_split' and 'max_depth'. To further optimize the performance of the Random Forest, an iteration was conducted over different values for the hyperparameters 'min_samples_leaf' and 'max_features', which are considered to be among the most crucial parameters. The results were plotted to enable the evaluation of the optimal value to be utilized for the model:

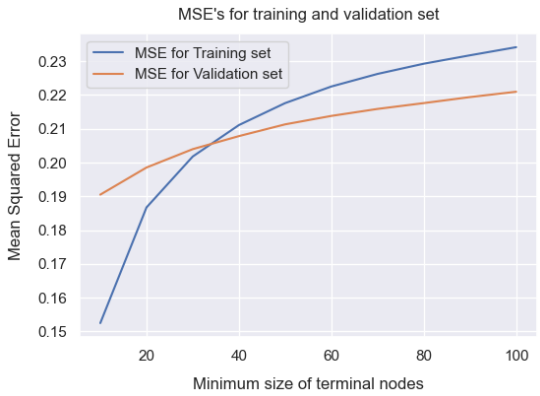

Figure 25: MSE for different min_samples_leaf

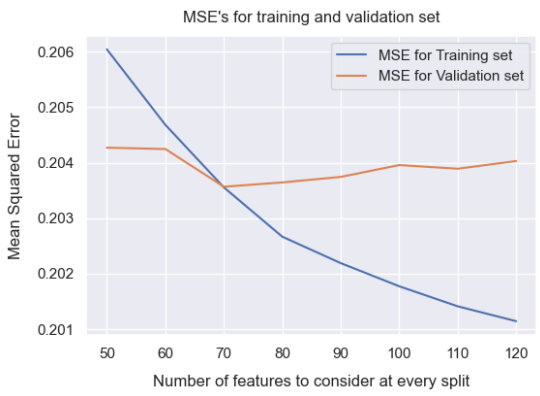

Figure 26: MSE for different max_features

The MSE in both samples (training and validation sets) is increasing for the minimum size of terminal nodes. However, the growth of the MSE is much faster in the training sample. A smaller MSE for a smaller minimum size of terminal nodes implies that the model is penalizing models that are more general or that lose details by not allowing a more significant minimum number of samples

to be at a leaf node. Smaller node size allows more detailed trees that can be better for prediction, despite the fact they can be more complex and can generate overfitting. We are choosing a minimum size of terminal nodes of 30 as, at that point, both MSE cross, so there is no trade-off. On the other hand, for the number of features considered at every split, the MSE for both training and validation samples is very low. The MSE for the train sample is decreasing with a higher number of features to consider at every split, while the MSE for the validation sample is relatively constant. We chose 70 features to consider at every split as it is where both MSE join and the lower for the validation set.

Based on this search for the most optimal parameters, it is decided that the selected model is the following:

$$RandomForestRegressor(n\_estimators = 700, max\_features = 70, max\_depth = 100,$$

$$min\_samples\_leaf = 30, min\_samples\_split = 10, random\_state = 0)$$

The MSE of the selected model was estimated on the test and validation sets and obtained results of 0.2035547 and 0.203566, respectively. Both MSEs are low and quite close, so it can be concluded that the selected model has good hyperparameters because it is neither overfitting nor underfitting. This selected model's prediction was also calculated on the test set and its respective MSE, which was 0.227295. Additionally, a histogram was made that allows us to see the similarity in the distribution of the observed and predicted index on the test set. It is evident that the predicted income index presents a distribution quite similar to the observed income index. However, the model captures above the actual index those individuals slightly above 1, which is the poverty threshold.

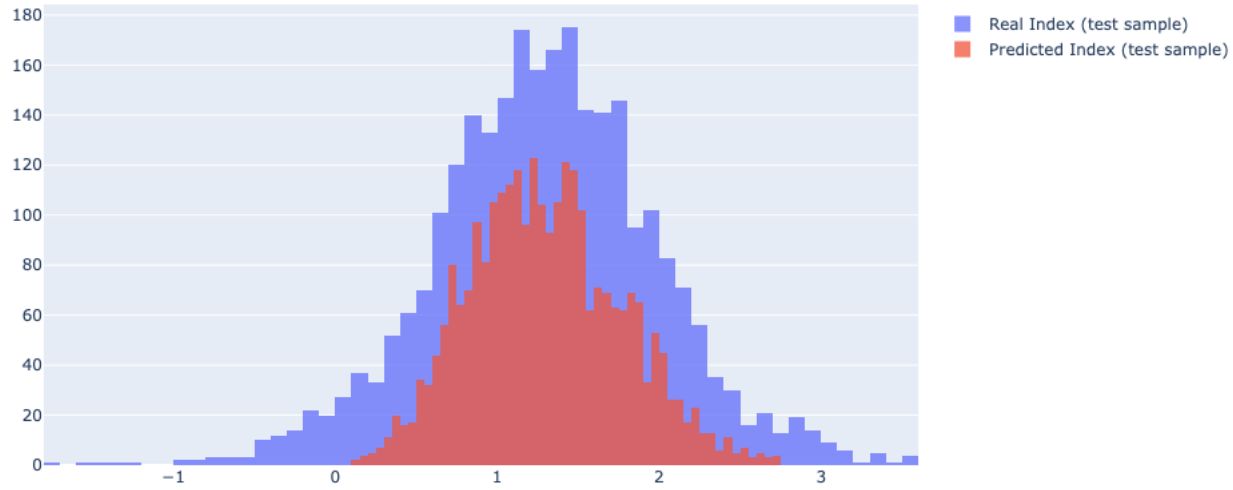

Figure 27: Predicted Distribution vs. Real distribution of the response variable in the test set

To analyze the quality of the predictions concerning the actual data, figures 28, 29, and 30 show the predicted income index on the y-axis and the real income index on the x-axis. The 45-degree line would represent a model with perfect predictive capacity. It is evident that the distribution of the training set, validation set, and test set is quite similar to the 45^{o} line, indicating that the model captures the data's behavior adequately. In order to show the exclusion and inclusion errors and based on the approximation used by Caballero (2021), the plane was divided into four quadrants, using two lines: the vertical represents the poverty threshold, and the horizontal represents the eligibility of access to the potential social program.

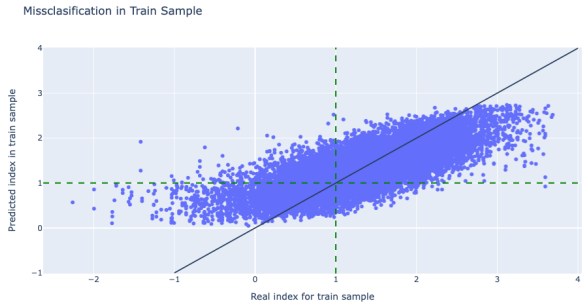

Figure 28: Model performance (Train sample)

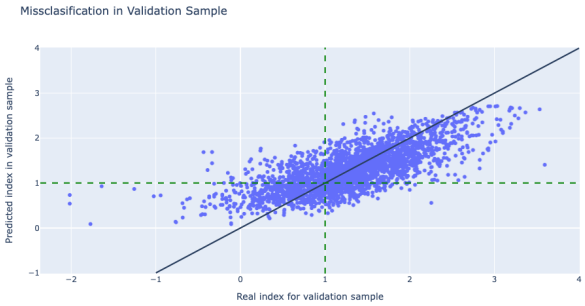

Figure 29: Model performance (Validation sample)

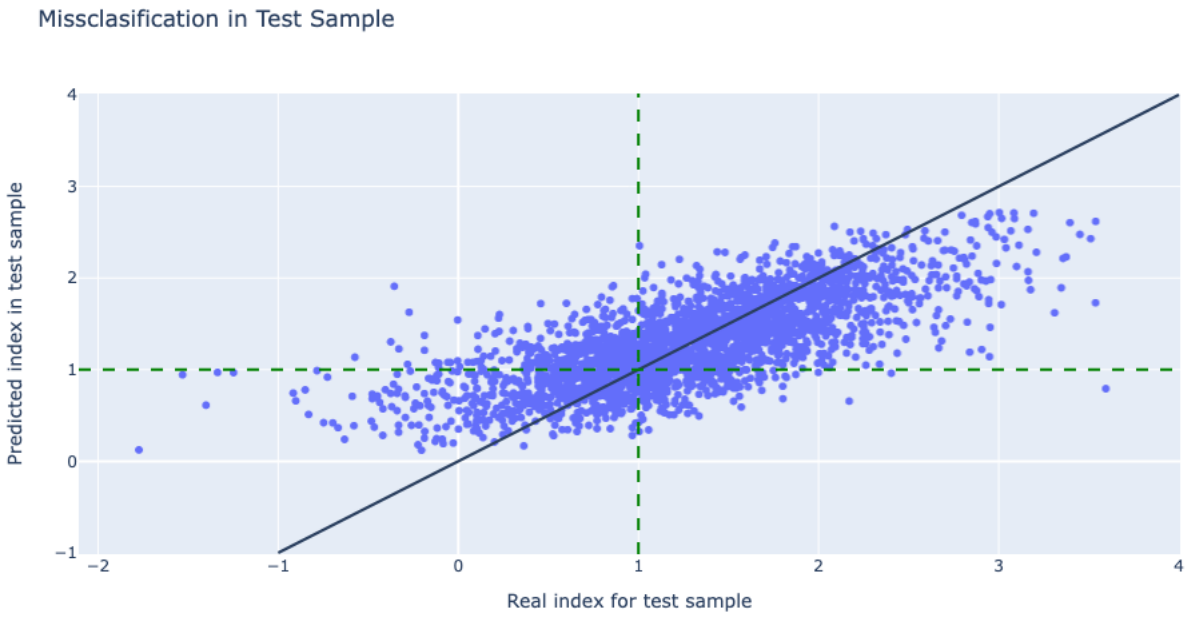

Figure 30: Model performance including inclusion and exclusion errors (Test sample)

The observations in the upper left quadrant can be considered exclusion errors since they are poor but are classified as non-poor under the model designed to give money transfers. Likewise, those in the lower right quadrant are observations that are part of the inclusion error since they are non-poor people but that the model and, therefore, the aid would be considered poor. To quantify the number of points that are misclassified, the confusion matrix for the test set was estimated:

	Poor	**Non-poor**
Predicted as poor	526	219
Predicted as non-poor	344	1611

3.2.3 Support Vector Machine

The second method we decided to implement for this prediction study is the Support Vector Machine (SVM). Although time-consuming when it is run, this model can be exceptionally performing and offers multiple options through hyperparameter tuning. As for the Random Forest model, the

initial dataset has been split as follows: 80% for the training set, 10% for the validation set, and 10% for the test set.

To run the SVM model, we also use the scaled data with StandardScaler, which removes the mean and divides by the variance. This is crucial because it allows the model to compare the variables equivalently without assigning too much weight to a variable that takes high values due to its nature.

The Support Vector Regressor (SVR) of scikit-learn proposes multiple parameters, including:

- Kernel: gives the kernel that will be used in the model. The kernel is the mathematical function that will take the data as input and provide an output that will be used to make predictions. The kernel can be linear or non-linear. We chose the second option because it is more adaptive to the data. We would have tried with a radial basis function (gaussian) kernel and a polynomial, but we restricted ourselves to an RBF for computing capacity reasons.
- C: cost parameter, makes a tradeoff between model accuracy and ability to predict unknown data. The higher C, the more permissive the model.
- Gamma: non-linear regressor parameter. The higher Gamma, the smaller the variance of the model.

First, we built an SVM model with an RBF kernel and default values for C and Gamma. The default value of C is 1, and for Gamma is 'scale', which corresponds to $1/(n_features * X.var()) \approx 0.0072$ for our data. To assess the quality of this first model, we compare the mean squared error on the training and test samples. We obtain, respectively, 0.111 and 0.174. We implemented a cross-validation method to find the best parameters for our SVM model using GridSearchCV. Again, because of running time, we had to restrict ourselves to a limited number of values taken by the hyperparameters. These are shown below.

Hyperparameter	**Values**
C	0.1, 1, 10, 100
Gamma	1, 0.1, 0.01, 0.001

The grid has been fit on the training datasets. This process gave the following optimized parameters: C = 10 and gamma = 0.001.

We hence built an improved SVM model with the parameters indicated above. Like before, we produce predictions on the train, validation, and test samples. The MSE gives the following results: on the training sample, the MSE has decreased by 36% with the improved model. On the validation sample, it is almost the same. The MSE on the test sample is 0.187, and the R-squared is 0.611. We will see in the next section that it outperforms the random forest model.

With this improved model, we obtain the following confusion matrix:

	Poor	**Non-poor**
Predicted as poor	557	185
Predicted as non-poor	313	1645

With this model, we get 24.93% of inclusion error (people that are classified as "poor" when they are not), and 35.98% of exclusion error (people that are classified as "non-poor" when they actually are poor).

In the same way as for the random forest, we built a graph to see the predicted index distribution compared to the actual index distribution. As for the previous model, we can see that the frequency of predicting an index in the middle of the distribution is higher than the real index, which has thicker tails on both sides.

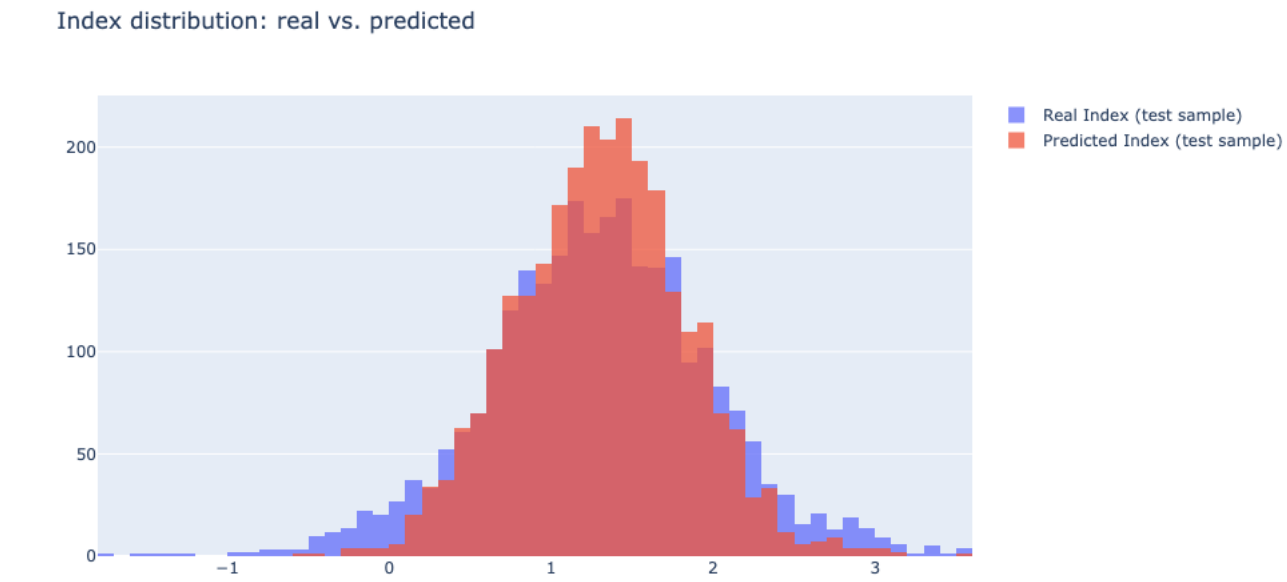

Figure 31: Predicted Distribution vs. Real distribution of the response variable in the test set

The following graphs were made with the same logic as for the random forest. The upper left quadrant can be interpreted as poor people predicted as non-poor, so excluded from an eventual cash transfer program.

Figure 32: Model performance (Train sample)

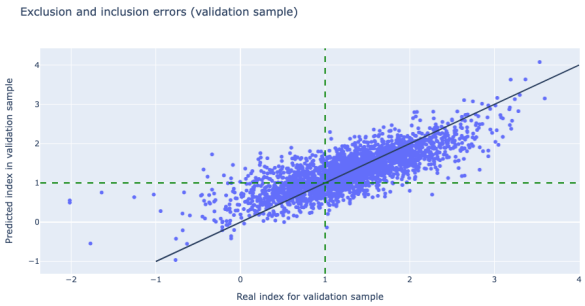

Figure 33: Model performance (Validation sample)

Figure 34: Predicted Distribution vs. Real distribution of the response variable in the test set

3.2.4 Model selection

The optimal model selection is based on a comprehensive evaluation of multiple performance metrics, including the Mean Squared Error (MSE), the coefficient of determination (R^2), and the confusion matrix. The confusion matrix, in particular, allows us to estimate the potential inclusion and exclusion errors that may occur if a conditional cash transfer program, based on the poverty threshold calculated using the index developed in this study, were to be implemented. The results of the performance metrics used to compare and select the best model can be found in the following table:

Table 6: Models comparison

	MSE (train set)	**MSE (validation set)**	**MSE (test set)**	R^2	**Exclusion Error**	**Inclusion Error**
Random Forest	0.2035547	0.2035662	0.2273	0.527	39.54%	29.4%
Support Vector Machine	0.152	0.173	0.187	0.611	35.98%	24.93%

After evaluating the results of these metrics, it was determined that the Support Vector Machine (SVM) was the optimal model, as it demonstrated lower MSEs and higher R^2 values than the Random Forest, indicating that it was able to better capture the underlying patterns in the data. Furthermore, the inclusion and exclusion errors were calculated based on the condition that the threshold for accessing the program was set at 1, in line with the established poverty threshold. The results of these evaluations are presented in table 6 and in figure 35. The equations used for computing them are the following:

$$Error_{Inclusion} = \frac{False\_Poor}{False\_poor + True\_poor}$$

$$Error_{Exclusion} = \frac{False\_non\_poor}{False\_non\_poor + True\_poor}$$

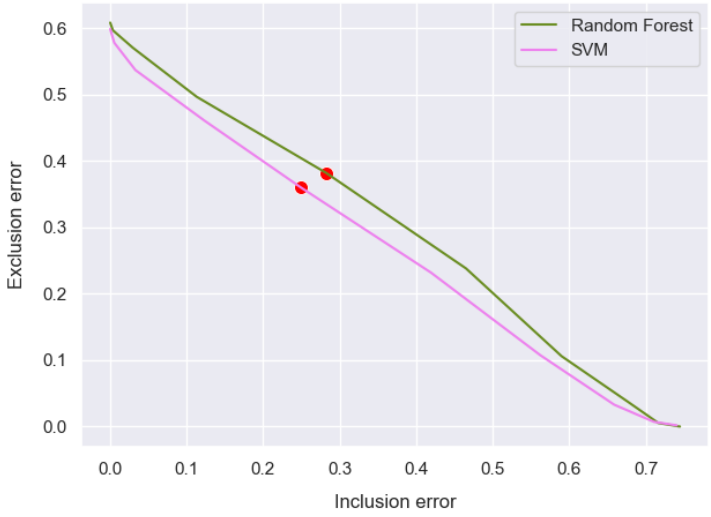

Figure 35: Inclusion and Exclusion error

As can be seen in the figure, the inclusion and exclusion errors are lower for the Support Vector Machine than for the Random Forest model in our study. Compared to the literature, the inclusion error of 24.9% for the SVM is lower than those reported in other studies, such as Caballero (2021) and Noriega-Campero (2020). The inclusion error in Colombia for 2019 was reported at 27.2% by Caballero (2021) and 30% for Colombia by Noriega-Campero (2020). On the other hand, the exclusion error for the SVM model is also low compared to the study by Caballero (2021), which reported an exclusion error of 44.3% for Colombia in 2019 when using the poverty line as the access threshold to the program. Nonetheless, it is slightly above the 30% reported by Noriega-Campero (2020). Overall, these results support the conclusion that the SVM model is the best model for this study.

In our analysis, the exclusion errors are higher than the inclusion errors for the Support Vector Machine, which indicates that some individuals who are truly poor may not be receiving aid under a poverty threshold established by the income index. The fact that the exclusion errors are higher than the inclusion errors is something that happens constantly in the literature that analyzes the effectiveness of poverty reduction policies. However, it also means that the government is allocating resources more efficiently compared to a situation of higher inclusion error, as they are not missing resources in the population that does not need it, that is, in people who are not really poor but who are considered as poor by the model. Nonetheless, there is room for improvement by reducing the exclusion error, which would enable the government to target more truly poor people.

It is important to note that the exclusion error may be higher in our analysis due to a number of factors such as the limitations of the data used, the complexity of the income index and the limitations of the model itself. Additionally, there may be other socio-economic factors that influence poverty and were not considered in the analysis, which could lead to a higher exclusion error. Despite these limitations, the results of the analysis can still be used to inform policy decisions and guide further research in the field.

3.2.5 Feature importance for Support Vector Machine

In order to understand which features contribute the most to the prediction in the SVM model, we attempted to use the SHAP values method. However, the computational capacity required for this method was beyond our means. As an alternative, we employed the scikit-learn's permutation feature importance technique. This method assesses the decrease in a model's performance when a single feature value is randomly altered. If the model's performance is significantly affected by the alteration, it is indicative of the feature's importance. The graph below illustrates the 10 most critical variables in the SVM model.

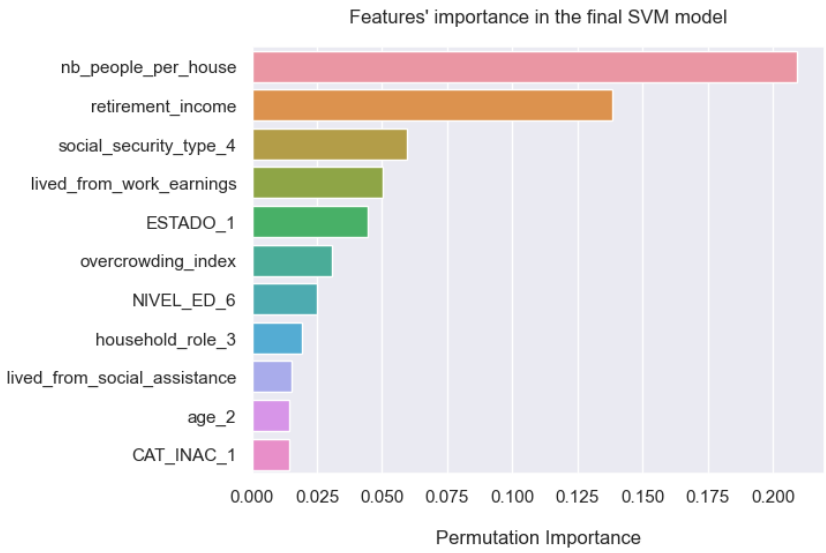

Figure 36: Feature importance for SVM

As seen in the graph, the family size and retirement income of the individual are the two most crucial variables in predicting the index. Additionally, other factors such as the type of social security, whether an individual's income is derived from work, and employment status also play a significant role in the prediction of the target variable. This information can be useful for future policy-making and program design to target specific groups and factors that are more likely to be living in poverty.

Since the computational cost for Random Forest is less, we will also show the SHAP values for the Random Forest Model to get more insights in the following section.

3.2.6 SHAP Values for Random Forest

SHAP values, or SHapley Additive exPlanations, explain the output of any machine learning model. They provide interpretable and locally accurate importance scores for each feature by computing the contribution of each variable to the prediction of a given instance. They are based on Shapley values from cooperative game theory and can be used to explain global and local forecasts.

In this study, we use SHAP values to provide interpretability and assess the impact of each variable on the prediction. This approach allows us to offer a comprehensive and useful analysis for policymakers, as they will better understand the most important factors in determining program eligibility.

We computed SHAP values using the shap library in Python. We first created an instance of the Explainer class, passing in the Random Forest model and the training data. Finally, we used the explainer to compute the SHAP values for the test data. In the following, we will present global and individual plots and the most valuable insights we can obtain for them. The global plots section contains three types: variable importance plot, cohort bar plot, and beeswarm summary plot. The individual plots will present a force plot, and a waterfall plot for two individuals predicted as poor and non-poor.

(a) Global plots

1. **Variable Importance Plot.** The Variable Importance Plot provides a visual representation of the relative importance of each feature in determining the income index. The plot displays the features in descending order of importance, measured by their average absolute SHAP values.

   The plot results indicate that the most important variable in determining the income index is social security type, with a value of +0.16, which suggests that the absence of health coverage is a major factor contributing to poverty. The second most important variable is one we have created,

the overcrowding index, with a value of +0.13, indicating that overcrowding in the household also plays a significant role in determining the income index.

Other important variables identified by the plot include the state of the individual, whether they are employed or not, with a value of +0.1; the number of people in the household, with a value of +0.07; and the level of education, with a value of +0.06. The fact that the individual lived from social assistance in the last three months also significantly impacts the predicted income index, with a value of +0.04. The plot also shows that retirement or pension income amount and living in the last three months of what they earn at work have a positive impact on the predicted income index, with values of +0.04 and +0.02, respectively, whereas if the fuel used for cooking was gas, has a positive effect with a value of +0.02.

These results can be used to identify potential areas for intervention and improvement. The plot suggests that access to health coverage, adequate housing, employment opportunities, and education are crucial in addressing poverty.

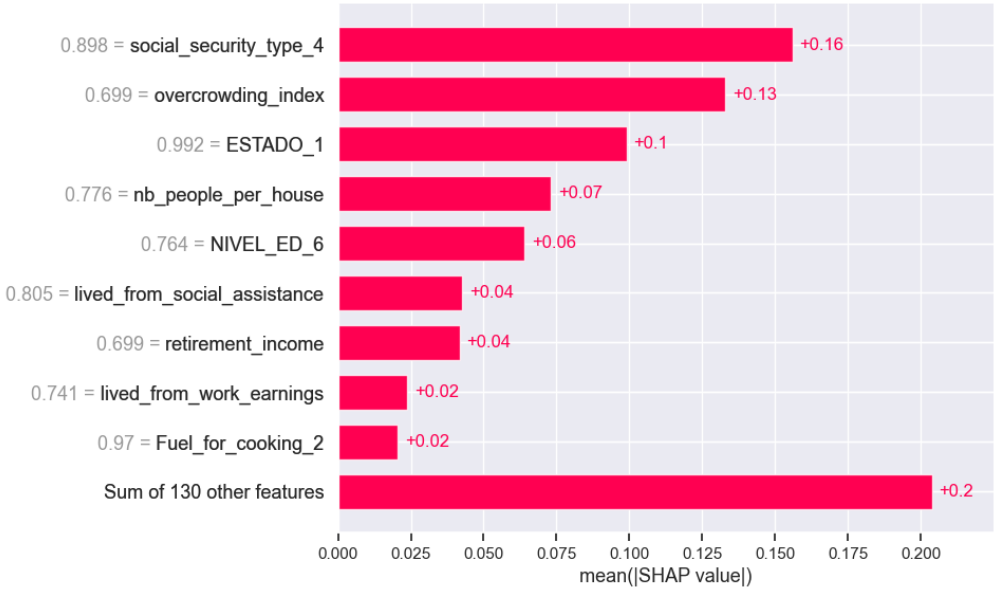

Figure 37: Variable Importance Plot

2. **Cohort Bar Plot.** The Cohort bar plot is a variation of the Variable Importance Plot, which allows for the comparison of the relative importance of each feature for different groups or cohorts. In this case, the cohorts are men and women, and the plot displays the average absolute SHAP values for each variable, separated by gender.

   The plot results indicate that the most crucial variable in determining the income index for both men and women is social security type, with values of +0.16 and +0.15, respectively. This suggests that the absence of health coverage significantly contributes to poverty for both genders. Also, the level of education is an important variable, with values of +0.06 for men and +0.07 for women. This indicates that educational attainment plays a significant role in determining the income index for both men and women.

   Interestingly, the variable "the person lived in the last three months of what they earn at work" has different values for men and women, with values of +0.02 and +0.03, respectively, which can imply that men and women have different financial situations. The Cohort bar plot provides a detailed view of the factors determining the income index and poverty for men and women. Furthermore, the graphs highlight the need to consider gender-specific factors when addressing poverty, as the impact of certain variables may differ between men and women.

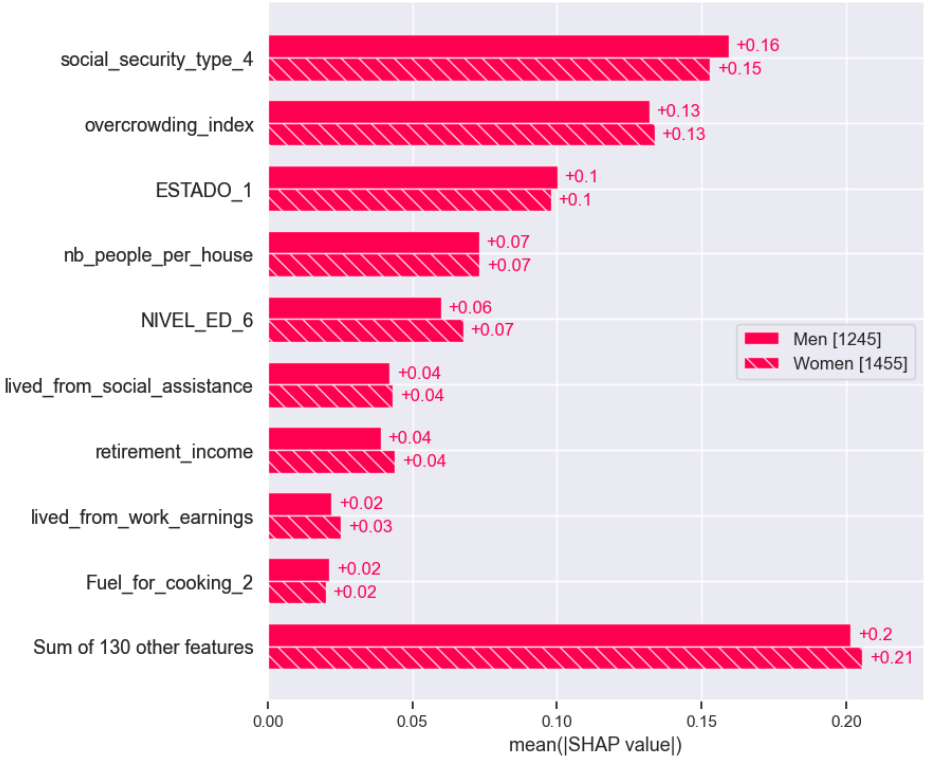

Figure 38: Cohort Bar Plot

3. **Beeswarm Summary Plot.** The Beeswarm Summary plot presents an information-dense summary of how the top features in a dataset impact the model's output. The y-axis displays the names of the variables, while the x-axis indicates the SHAP values. The plot uses a color coding system, where red represents high feature values, and blue represents low feature values. It reveals, for example, that a high retirement income increases the predicted income index.

   The results of the Beeswarm summary plot align well with the reality and the characteristics of what one would expect from someone who is considered poor or not poor. Factors such as having a higher retirement or pension income amount, being employed, and having a full university degree completed, are commonly associated with higher income and a higher income index. These individuals are more likely to have financial stability and better job opportunities, which can contribute to a higher income index.

   On the other hand, factors such as not having social security, living from social assistance, living in the last three months from merchandise, clothing, or food from the government, churches, or schools, having a higher overcrowding index, and a higher number of people in the house, are commonly associated with lower income and a lower income index. These individuals are more likely to lack access to basic necessities such as health coverage and food security and may be more prone to poverty.

   The Beeswarm summary plot aligns well with reality and provides insights into the factors that contribute to poverty. It can identify potential areas for intervention and improvement and inform policies and programs aimed at addressing poverty.

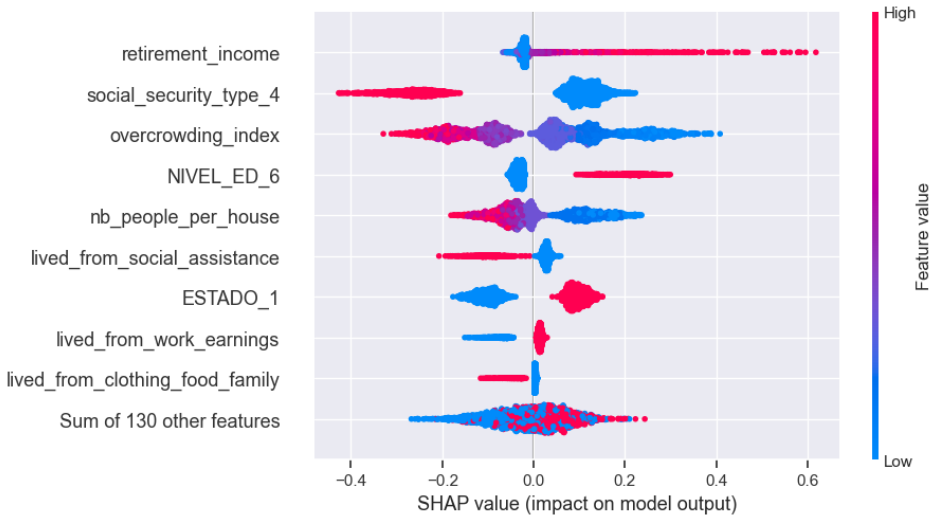

Figure 39: Beeswarm Summary Plot

(b) Individual plots

The Force plot and Waterfall plot are visual representations of which features most influenced the model's prediction for a single individual. They show the importance of considering individual-specific factors when addressing poverty, as the impact of certain variables may differ from one individual to another. In both plots, the essential features to making the prediction are shown in red and blue, with red representing features that pushed the model score higher and blue illustrating features that moved the score lower.

The Force plot shows the features that had a higher impact on the score closer to the dividing boundary between red and blue, and the bar represents the size of that impact. On the other hand, the Waterfall plot shows the feature's importance as a stacked bar chart. It provides an easy way to identify each feature's exact contribution score by showing its cumulative effect on the final prediction.

For a poor and a non-poor individual, we plot both the Force plot and Waterfall plot, as they complement each other by providing a detailed view of the feature's importance. The Force plot allows an understanding of how each feature value impacts the final prediction. In contrast, the Waterfall plot provides an easy way to identify the exact contribution score of each feature. Together, these two plots offer a comprehensive view of the feature importance for a specific case and allow the identification of critical factors driving the prediction.

1. **Individual 1 (non-poor).** The plot results indicate that this individual's initial prediction is 1.25. However, after considering the feature values, the final prediction is 2.1, meaning that this individual is predicted as non-poor. This suggests that some feature values positively impact the predicted income index and push the forecast higher.

   The plot shows that the level of education of a university degree completed is the most essential feature, pushing the prediction higher by +0.27. This aligns with the reality and the characteristics of what one would expect from someone with higher education, as they are more likely to have access to better job opportunities and a higher income.

   Also, a low number of people in the house, having social security, being employed, having a low overcrowding index, and not living from social assistance are other vital features that have a positive contribution on the final predicted income index, with values of +0.19, +0.17, +0.12, +0.08, +0.03 and +0.02, respectively.

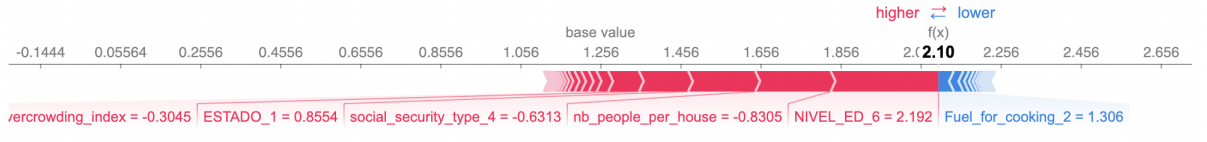

Figure 40: Force Plot for non-poor individual

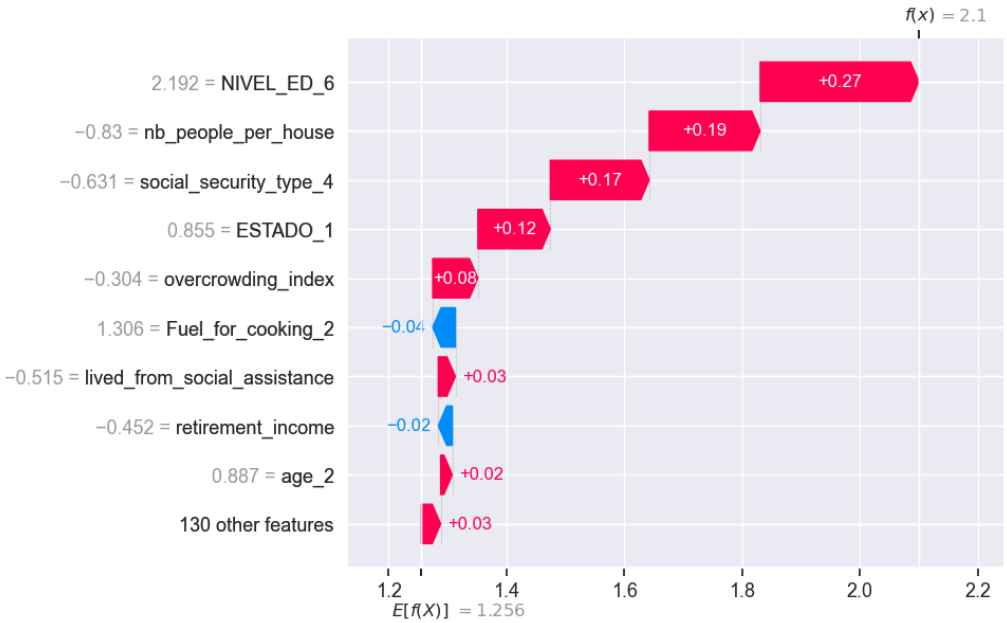

Figure 41: Waterfall Plot for non-poor individual

2. **Individual 2 (poor).**

   The baseline prediction for this individual is 1.25, but after considering the feature values, the final prediction is 0.62. For this individual predicted as poor, these plots highlight the key factors that determine their income index, which can be used to identify potential areas for intervention and improvement for this individual.

   The plot indicates that the most critical features are not having social security, a high overcrowding index, living from social assistance, an increased number of people in the house, and failing to achieve a university degree. They have a negative contribution to the final predicted income index, with values of -0.22, -0.21, -0.14, -0.08, and -0.02, respectively.

   On the other hand, having a laundry machine and being employed are the most important features that positively contribute to the final predicted income index, with values of +0.02 and +0.1, respectively. This suggests that even though these factors may not be enough to lift an individual out of poverty, they could help improve the individual's living conditions and financial situation.

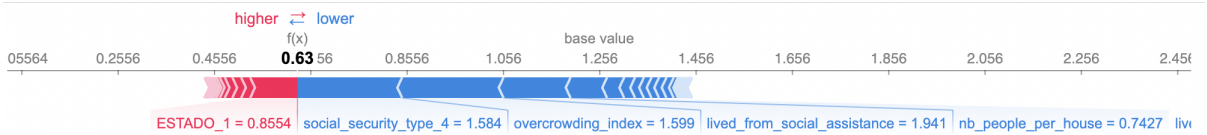

Figure 42: Force Plot for poor individual

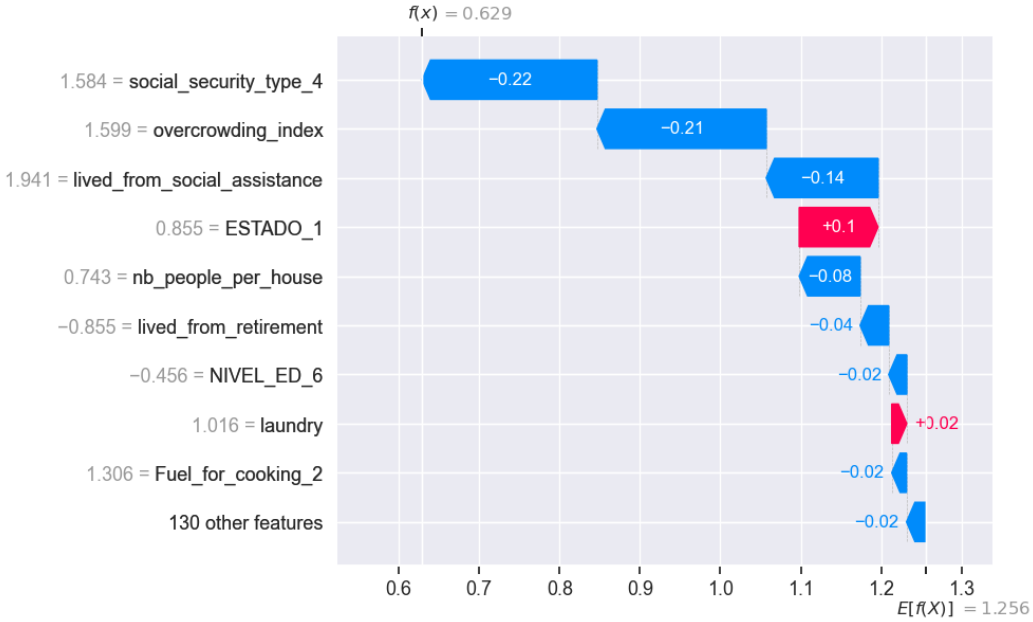

Figure 43: Waterfall Plot for poor individual

4 Conclusion

In this study, we aimed to evaluate the performance of various machine learning models to predict the income of Argentinians. The data used for this analysis included information on individuals' demographic characteristics, education levels, sources of income, and housing characteristics.

Our findings showed that the Support Vector Machine (SVM) model performed best in prediction accuracy, as it had the lowest mean squared error (MSE) and the highest R-squared value compared to the Random Forest model. Additionally, the inclusion and exclusion errors were lower for the SVM model, indicating that it was more effective in identifying impoverished individuals.

Furthermore, the analysis of SHAP values revealed that health coverage, adequate housing, employment opportunities, and education were the most important predictors of the income index. We also illustrated the features that most influenced the model prediction for a poor and a non-poor individual, allowing us to identify the different drivers in both cases.

Regarding limitations, the exclusion error was higher than the inclusion error. Despite this, it is essential to note that this fact occurs in other studies. Efforts can be made to improve the accuracy and precision of the models and data collection methods in future studies to reduce inclusion and exclusion errors.

As for possible future work, one possibility would be to incorporate additional variables or try another dimensionality reduction approach into the analysis to improve the accuracy of predictions. Another option would be to conduct further testing of different machine learning models to see if any other model outperforms Support Vector Machine, like XGBoost or gradient boosting, which are widely used in the literature. Additionally, the study could be extended to other countries with similar characteristics, such as high rates of informality and poverty, to test the generalizability of the findings.

5 References

Becker, G. S., & Chiswick, B. R. (1966). Education and the Distribution of Earnings. The American Economic Review, 56(1/2), 358–369. http://www.jstor.org/stable/1821299

Bonaglia, F. (2019). Programas de transferencias condicionadas en la región, Mecanismos de focalización desde una perspectiva comparada entre Uruguay y Costa Rica. Universidad de la República, Facultad de Ciencias Sociales, Departamento de Sociología.

Chen, T. (2018). Evaluating Conditional Cash Transfer Policies with Machine Learning Methods. Washington University in St. Louis. https://doi.org/10.48550/arXiv.1803.06401

ECB Statistics Paper No 15 (2016). Statistics Paper Series. Unit non-response in household wealth surveys. European Central Bank.

Galvis Caballero, A. (2021). ¿Cómo puede contribuir el machine learning a la focalización de programas sociales? Universidad Autónoma de Bucaramanga, Bucaramanga, Colombia.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R. Springer Texts in Statistics. ISBN 978-1-4614-7137-0, doi: 10.1007/978-1-4614-7138-7

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

Matkowski, J. (2021). Prediction of Individual Level Income: A Machine Learning Approach. Bryant University Honors Program.

Noriega-Campero, A., Garcia-Bulle, B., Cantu, L.F., Bakker, M.A., Tejerina, L., & Pentland, A. (2020). Algorithmic Targeting of Social Policies: Fairness, Accuracy, and Distributed Governance. In Conference on Fairness, Accountability, and Transparency. January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3351095.3375784

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 41, 647–665. https://doi.org/10.1007/s10115-013-0679-x

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 67(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

6 Appendix

People's characteristics		
Variable in our model	**Variable in the EPH**	**Description**
NIVEL_ED	NIVEL_ED	1 = Incomplete primary (includes special education) 2 = complete primary 3 = Incomplete high school 4 = full secondary 5 = Incomplete university degree 6 = Full university degree 7 = No instruction
literate	CH09	Can you read and write? 1 = Yes 2 = No
marital_status	CH07	1 = joined 2 = married 3 = separated or divorced 4 = widow/widower 5 = single
household_role	CH03	1 = Boss 2 = Spouse/partner 3 = Son/stepson 4 = Son-in-law / daughter-in-law 5 = Grandchild 6 = Mother/father 7 = Father-in-law 8 = Sibling 9 = Other relatives 10 = Not relatives
sex	CH04	1 = male 2 = female
age	CH06	Discrete variable
REGION	REGION	Region code 1 = Gran Buenos Aires 40 = NOA 41 = NAE 42 = Cuyo 43 = Pampeana 44 = Patagonia

Variable in our model	Variable in the EPH	Description
AGLOMERADO	AGLOMERADO	Where he/she lives (agglomerate code): 2 = Gran La Plata 3 = Bahía Blanca - Cerri 4 = Gran Rosario 5 = Gran Santa Fé 6 = Gran Paraná 7 = Posadas 8 = Gran Resistencia 9 = Comodoro Rivadavia - Rada Tilly 10 = Gran Mendoza 12 = Corrientes 13 = Gran Córdoba 14 = Concordia 15 = Formosa 17 = Neuquén – Plottier 18 = Santiago del Estero - La Banda 19 = Jujuy-Palpalá 20 = Río Gallegos 22 = Gran Catamarca 23 = Gran Salta 25 = La Rioja 26 = Gran San Luis 27 = Gran San Juan 29 = Gran Tucumán - Tafí Viejo 30 = Santa Rosa – Toay 31 = Ushuaia - Río Grande 32 = Ciudad Autónoma de Buenos Aires 33 = Partidos del GBA 34 = Mar del Plata 36 = Río Cuarto 38 = San Nicolás – Villa Constitución 91 = Rawson – Trelew 93 = Viedma – Carmen de Patagones

Housing characteristics		
Variable in our model	**Variable in the EPH**	**Description**
floor_type	IV3	The interior floors are mainly made of... 1 = mosaic/tile/wood/ceramic/carpet 2 = fixed cement/brick 3 = loose brick/dirt 4 = others
roof_type	IV4	The outer roof covering is... 1 = membrane/asphalt cover 2 = tile/slab without cover 9 = Don't know. Apartment in horizontal property
housing_type	IV1	Housing type: 1 = home 2 = department 3 = tenancy room 4 = room in hotel/pension 5 = local not built for habitation 6 = other
kitchen	II4_1	kitchen room 1 = Yes 2 = No
laundry	II4_2	laundry room 1 = Yes 2 = No
garage	II4_3	garage 1 = Yes 2 = No
room_to_sleep	II5	Do you use any of the rooms for sleeping?(rooms previous questions) 1 = Yes 2 = No
room_to_work	II6	Some of the rooms are used exclusively as a workplace: office, studio, workshop, business, etc.(rooms previous questions) 1 = Yes 2 = No
Fuel_for_cooking	II8	Fuel used for cooking: 1 = Mains gas 2 = Gas pipe/bottle 3 = Kerosene/firewood/coal 4 = other
bathroom	II9	Bathroom (possession and use): 1 = Exclusive use of the home 2 = Shared with other household(s) from the same dwelling 3 = Shared with another home/s 4 = Does not have a bathroom

Variable in our model	**Variable in the EPH**	**Description**
sewerage	IV11	The bathroom drain is... 1 = to the public network (sewer) 2 = to septic chamber and cesspool 3 = cesspool only 4 = to hole/excavation in the ground
access_to_water	IV6	It has water... 1 = by pipe inside the house 2 = outside the house but inside the lot 3 = out of bounds
'trash_nearby'	IV12_1	The dwelling is located near a garbage dump/es (3 blocks or less) 1 = Yes 2 = No
'trash_nearby'	IV12_2	The house is in a flood zone (in the last 12 months) 1 = Yes 2 = No
shantytown	IV12_3	The house is in a shantytown (by observation) 1 = Yes 2 = No
nb_people_per_house	IX_TOT	Discrete variable
nb_rooms	II1	Discrete variable
overcrowding_index	IX_TOT/II1	Persons / Rooms

Socioeconomic characteristics		
Variable in our model	**Variable in the EPH**	**Description**
ESTADO	ESTADO	Activity Condition: 1 = Employed 2 = Unemployed 3 = Inactive
CAT_OCUP	CAT_OCUP	Occupational category (with a previous occupation): 1 = Boss 2 = Own account worker 3 = Laborer or employee 4 = Unpaid family worker
CAT_INAC	CAT_INAC	Inactivity category: 1 = Retired / Pensioner 2 = Rentier 3 = Student 4 = Housewife 6 = Disabled 7 = Other
social_security_type	CH08	Do you have any type of health coverage that you pay for or are discounted? 1 = Social work (includes PAMI) 2 = Mutual / prepaid / emergency service 3 = Public plans and insurance 4 = Does not pay nor is discounted 9 = Don't know/No response 12 = Social and mutual work / prepaid / emergency service 13 = Social work and public plans and insurance 23 = Mutual / prepaid / emergency service / Public plans and insurance 123 = Social work, mutual / prepaid /emergency service and public plans and insurance
property_ownership_regime	II7	Housing tenure regime 1 = Owner of the house and the land 2 = Homeowner only 3 = Tenant / lessee of the dwelling 4 = Occupant for payment of taxes/expenses 5 = Occupant in a dependency relationship 6 = Free occupant (with permission) 7 = Actual occupant (without permission) 8 = Is in succession 9 = Other

Main Income		
Variable in our model	**Variable in the EPH**	**Description**
lived_from_work_earnings	V1	Living in the last 3 months: ...of what they earn at work? 1 = Yes 2 = No
lived_from_retirement	V2	Living in the last 3 months: ...of any retirement or pension? 1 = Yes 2 = No
lived_from_severance_pay	V3	Living in the last 3 months: ...of severance pay? 1 = Yes 2 = No
lived_from_ unemployment_insurance	V4	Living in the last 3 months: ...unemployment insurance? 1 = Yes 2 = No
lived_from_social_assistance	V5	Living in the last 3 months: ...subsidy or social aid (in money) from the government, churches, etc.? 1 = Yes 2 = No
lived_from_clothing_food_gov	V6	Living in the last 3 months: ...with merchandise, clothing, government food, churches, schools, etc.? 1 = Yes 2 = No
lived_from_clothing_ food_family	V7	Living in the last 3 months: ...with merchandise, clothing, food from relatives, neighbors, or other people who do not live in this household? 1 = Yes 2 = No
lived_from_rent_own_property	V8	Living in the last 3 months: ...any rent (for a house, land, office, etc.) of your property? 1 = Yes 2 = No
lived_from_business_earnings	V9	Living in the last 3 months: ...profits from a business they don't work for? 1 = Yes 2 = No
lived_from_investments	V10	Living in the last 3 months: ...interest or income for fixed terms / investments? 1 = Yes 2 = No

Variable in our model	**Variable in the EPH**	**Description**
lived_from_scholarship	V11	Living in the last 3 months: ...a scholarship 1 = Yes 2 = No
lived_from_cash_family	V12	Living in the last 3 months: ...food allowances or cash assistance from people who do not live in the household? 1 = Yes 2 = No
lived_from_savings	V13	Living in the last 3 months: ...spend what they had saved? 1 = Yes 2 = No
lived_from_loans_family	V14	Living in the last 3 months: ...ask family/friends for loans? 1 = Yes 2 = No
lived_from_loans_bank	V15	Living in the last 3 months: ...ask for loans from banks, finance companies, etc.? 1 = Yes 2 = No
rent_own_property_income	V8_M	Amount of rental income (home, land, kitchen, etc.) from your property
aid_family_income	V12_M	Amount of income from food quotas or cash assistance from people who do not live in the home
investments_income	V10_M	Amount of income from interest or income for fixed terms/investments
retirement_income	V2_M	Retirement or pension income amount
government_aid_income	V5_M	Amount of income for subsidy or social assistance (in money) from the government, churches, etc.