Machine learning approach for multidimensional poverty estimation

Enfoque de aprendizaje automático para la estimación de la pobreza multidimensional

Mario Ochoa[1] https://orcid.org/0000-0002-2283-0203, Ricardo Castro-García[2] https://orcid.org/0000-0002-2332-9381, Alexander Arias Pallaroso[1,3] https://orcid.org/0000-0002-6612-4968, Antonia Machado[1] https://orcid.org/0000-0003-3856-5659, Dolores Sucozhañay Machado[4] https://orcid.org/0000-0003-4930-2176

[1] *Grupo de Investigación en Población y Desarrollo Local Sustentable (PYDLOS), Departamento Interdisciplinario de Espacio y Población (DIEP), Universidad de Cuenca,* Cuenca, Ecuador
mario.esteban.ochoa@gmail.com, antonia.machado@ucuenca.edu.ec

[2]*Microsoft*
ricardo.castro@microsoft.com

[3] *Facultad de Jurisprudencia y Ciencias Políticas y Sociales, Universidad de Cuenca,* Cuenca, Ecuador.
alexander.arias@ucuenca.edu.ec

[4] *Departamento Interdisciplinario de Espacio y Población (DIEP), Facultad de Ciencias Económicas y Administrativas, Universidad de Cuenca,* Cuenca, Ecuador.
dolores.sucozhanay@ucuenca.edu.ec

Resumen

En las ciencias sociales ha predominado un análisis teórico en sus investigaciones. La escasez de datos, su dificultad para recolectarlos y almacenarlos, ha sido la principal limitación para que las ciencias sociales adopten enfoques cuantitativos. Sin embargo, la gran cantidad de información generada en los últimos años, principalmente a través del uso de Internet, ha permitido que las ciencias sociales incluyan cada vez más análisis cuantitativos. Este estudio propone el uso de tecnologías como Machine Learning (ML) para solventar esta escasez de datos. El objetivo es estimar el índice de pobreza multidimensional a nivel de persona en un territorio en particular de Ecuador mediante el uso de modelos de regresión de Machine Learning (ML) partiendo de una cantidad limitada de datos para entrenamiento. Se comparan

Sumario: Introduction, Related work, Methodology, Results, Discussion and Conclusions.

Como citar: Ochoa, M., Castro-García, R., Arias, A., Machado, A. & Sucozhañay, D. (2021). Machine Learning Approach for Multidimensional Poverty Estimation. *Revista Tecnológica - Espol, 33(2), 205-225.* http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/853

10 modelos ML, tales como modelos lineales, regularizados y ensamblados. Random Forest se desempeña de manera sobresaliente frente a los otros modelos. Se llega a un error del 7,5% en la validación cruzada y del 7,48% con el conjunto de datos de prueba. Las estimaciones se comparan con aproximaciones estadísticas del IPM en una zona geográfica y se obtiene que el IPM promedio estimado por el modelo en comparación con el promedio informado por los estudios estadísticos difiere en 1%.

Palabras clave: bosques aleatorios, ciencias sociales, regresión, zona geográfica, datos limitados.

Abstract

In the social sciences, a theoretical analysis has predominated in its research. The scarcity of data and its difficulty in collecting and storing it, has been the main limitation for the social sciences to adopt quantitative approaches. However, the large amount of information generated in recent years, mainly through the use of the Internet, has allowed the social sciences to include more and more quantitative analysis. This study proposes the use of technologies such as Machine Learning (ML) are the answers to solving this data scarcity. The objective is to estimate the multidimensional poverty index at the personal level in a particular territory of Ecuador by using Machine Learning (ML) regression models based on a limited amount of data for training. Ten ML models are compared, such as linear, regularized, and assembled models and Random Forest performs outstandingly against the other models. An error of 7.5% was obtained in the cross-validation and 7.48% with the test data set. The estimates are compared with statistical approximations of the MPI in a geographical area and it is obtained that the average MPI estimated by the model compared to the average reported by the statistical studies differs by 1%.

Keywords: random forest, social sciences, regression, region, limited dataset.

Introduction

In recent years, the social sciences have increasingly chosen to use inductive analysis methodologies in their studies. Contrary to the trend over the years to use deductive approaches by default (Grimmer et al., 2021). In the past, most sociological studies adopted the deductive approach due to the recurring problem in the social sciences, the scarcity of data (Chen et al., 2018; Grimmer et al., 2021). Before reviewing or collecting data, the social scientist usually had a clear theory from which to draw propositions that could be testable. From these propositions, the variables of interest were raised and strategies were developed for their measurement, to finally establish hypotheses and a research design that tested the validity of the theory through the analysis of experimental observations (King et al., 1995; Rudin, 2015. This deductive approach made the researchers miss the opportunity to redefine their concepts, develop new theories and outline new hypotheses (Grimmer et al., 2021). In those years, data for social inductive analyzes was difficult to obtain; conducting surveys were expensive and storing large amounts of data was almost impossible. At the same time, the computing capacity was limited (Grimmer et al., 2021; Maldonado, 2019). Nowadays, the social sciences show a strong trend towards quantitative research. Sociologists have made many efforts to include data analysis in their research through statistical techniques such as descriptive statistics (tendency and dispersion analysis), inferential statistics (estimation of confidential interval and significance testing), and regression analysis (predictions and forecasting) (Rovai et al., 2013).

The advent of the Internet and the increase in computational capacity has undoubtedly made the amount of data go from being scarce to being immensely abundant. In this panorama, the social sciences have had to evolve to adapt to this new reality. The social sciences have

moved from concepts of "variables" to the "big data" idea. The methodologies based on observation, description, hypothesis formulation have become insufficient to analyze the current complex reality, driven by data and permanent virtuality (Maldonado, 2019).

This abundant amount of data has created the opportunity for the social sciences to move from the deductive approach to an inductive and iterative approach. Allowing to test hypotheses in a more agile way. Similarly, social networks allow creating studies and social experiments that until a few years ago were logistically impossible (Chen et al., 2018; Grimmer et al., 2021). The development of new and better data analysis algorithms has enabled social science studies to be able to establish more accurate estimations of social phenomena (Hindman, 2015).

In recent years, technologies that take advantage of all the possibilities derived from this vast amount of available data have been developed. As part of these advances, Data Mining (DM) has been one of the main technologies that benefited most.

DM is the science that focuses on extracting information from large data sets through the use of techniques from different disciplines such as ML or statistics (Hand, 2007). On the other hand, ML is a field of study that mainly develops algorithms and techniques to build models based on data. Depending on the structure of the data and the techniques used, ML is broadly divided into supervised learning, unsupervised learning, and reinforced learning (Chen et al., 2018). Its applications increasingly have been adopted by fields like economics, political science, and sociology (Molina & Garip, 2019).

ML has gained popularity and has great application potential in some social science disciplines. Its algorithms have been applied in tasks such as pattern prediction, characterization of population heterogeneity, causal inference, theory development, and support in experimental decision-making (Chen et al., 2018; Grimmer, 2015; Grimmer et al., 2021; Molina & Garip, 2019).

Despite the great benefits of using modern data analysis in social studies. The adoption of these techniques has been relatively slow compared to other fields (Lazer et al., 2009). Researchers have argued that the "black box" nature of the models used in ML, which does not allow knowing the internal inference process of the model, is one of the main factors for its rejection (Hindman, 2015). On the other hand, the lack of data from minority sectors of the population can cause models to be trained with a significant bias towards stereotypes, discrimination, and even racism (Chen et al., 2018). Another barrier that must be overcome is the lack of experience of data scientists in topics related to causal inference and complex social data processing (Grimmer, 2015).

However, social scientists could contribute with their expertise in human behavior, to establish a stronger connection between social science and ML to develop applications and tools focused on data generated by social studies (Chen et al., 2018).

As part of this effort to bridge the gap between these two fields of study, this paper uses ML techniques to improve analyzes of poverty. Our case study focuses on the multidimensional poverty of the city of Cuenca, one of the main cities in Ecuador. However, the same methodology can be applied to all the cities in the country.

The main objective is to create a model capable of estimating the MPI for each person living in the city, starting from a limited amount of data available for training. The task of the

model is to establish relationships between sociodemographic characteristics of people to estimate the corresponding MPI. In this sense, the research question is: What machine learning model allows the most accurate estimation of multidimensional poverty at the person level using predictive variables of a sociodemographic nature and with a limited amount of training data?

The databases used in this study corresponds to the National Survey of Employment, Unemployment, and Underemployment (ENEMDU) and Census of Population and Housing (CPH). The ENEMDU database contains sociodemographic variables and the calculations to obtain the MPI. However, the data in this database is limited because the survey is carried out following a probabilistic sampling strategy. This database is used for the training and validation process. On the other hand, the CPH database has observations for all the populations but it does not have enough variables to calculate de MPI. The model will be applied to this database to obtain an estimation of the MPI for all the population.

In section 2, the main concepts and related works about social science, ML, and multidimensional poverty are described. Section 3 explains in detail the followed methodology, including the data preprocessing in section 3.A, the model exploration in section 3.B, and fine-tuning in section 3.C. Following, section 4 presents the results obtained from the application of the models in the test dataset and the CPH database. A discussion about the obtained results is carried out in section 5. In section 6 some conclusions and future work are stated.

Related work
Social Science and Machine Learning

It is important to find a common point where the knowledge of ML and Social Science can be harnessed. However, before finding this common point, it is important to analyze the differences and similarities between these fields to have a better perspective of what they can contribute to the other. First, the philosophy behind the analysis in these fields is very different. Quantitative sociological studies begin with the proposition of research questions and hypotheses about how the world works and through theorizations and assumptions around the available data, it tries to explain a social phenomenon or individual and collective behaviors (Rudin, 2015). On the other hand, in ML the objective is not to explain the phenomena but to predict an outcome based on the only premise that the data has been obtained independently from an unknown distribution. It can then be said that sociological theory is hypotheses-driven, while ML is data-driven. Generally, ML starts with the data to later create a hypothesis, while sociology starts with the hypothesis statement from the beginning (Chen et al., 2018; Rudin, 2015).

Another important difference is the way the models are handled. In quantitative studies in sociology, the model (usually a linear regressor) is theorized and established before analyzing the data. However, in ML, several models are tested to determine which one best fits the data. Since ML focuses mainly on prediction for unseen data, its models do not provide a causal explanation for the analyzed phenomena, but rather emphasize adjusting the models and improving the accuracy of the predictions. This is a problem from a sociological point of view where the objective is to explain the whys and hows of observed phenomena (Chen et al., 2018; Rudin, 2015).

In addition, sociology generally generates conclusions based on multiple sources of information. Meanwhile, ML only works with a single data source. That is why applying ML methodologies in social analysis is a great challenge for researchers (Chen et al., 2018; Rudin, 2015). However, recent studies in the field of ML have focused their efforts on explaining the

phenomena, and although there is still only one source of data, the technical challenges have been raised to be able to analyze multiple sources to be coupled with the methodologies of the social sciences (Wallach, 2016).

One of the fields of study of the social sciences, and the one that concerns this study, is poverty. The notion of poverty is based on a value judgment regarding what are the minimally adequate levels of well-being in societies; it also refers to the degrees of deprivation that are intolerable (Moreno, 2017). There are three approaches to measuring poverty: the first corresponds to the utilitarian approach in which poverty is associated with the disposition of monetary income; a second approach is linked to Rawls's approach to justice, who raises the notion that societies must guarantee access to a series of primary goods or basic needs; lastly, there is the capabilities approach whose conceptual basis is articulated around poverty, human freedoms and the multidimensional nature of poverty (Denis et al., 2010).

Currently, poverty is estimated through socioeconomic household surveys. This approach is costly and time-consuming to carry out (Devarajan, 2013). However, below are described several studies that make use of ML techniques to enhance the estimation of poverty.

In Talingdan (2019) different machine learning algorithms are compared for the classification between a poor and a non-poor person based on information from the Community Based Monitoring System (CBMS) in Lagangilang, Abra, Philippines. Information about Health, Nutrition, Housing, Water and sanitary systems, education, income, employment, peace, and order, is considered. Algorithms such as Naive Bayes (NB), ID3, Decision Trees, Logistic Regressions, and K-Nearest Neighbors (KNN) are used. Naïve Bayes classifier outperformed all the four algorithms for predicting households that are poor and non-poor.

Kshirsagar et al. (2017) apply machine learning techniques to generate ProxyMeans Tests (PMT) which are quick surveys that allow estimating the probability that a person is poor or not, based on an approximation through the selection of variables related to poverty. The database used is based on national household surveys in Zambia. A set of variables is selected, a model is estimated using those variables to predict the poverty level of a household. The model obtained is capable of classifying poor households from non-poor households and is also applicable to different population segments of the country.

Similarly, Kambuya (2020) applies the Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest (RF) models, to improve the variable selection process and the performance of the PMT models. The results show that RF-based PMTs for their selection of variables, reduce the number of poor households classified as non-poor (exclusion error) and increase the accuracy when estimating the poverty rate at the national, urban, and rural levels. While the selection of variables based on LASSO, present better results compared to RF in terms of reducing the inclusion error (non-poor households classified as poor). Sohnesen & Stender (2017) also, use LASSO and RF to predict poverty using data of a period of one year. The results indicate that RF is a good predictor of poverty and obtains more robust estimates compared to Linear Regression estimators.

Several studies conclude that RF is one of the best algorithms for estimating poverty. Otok & Seftiana (2014) determines that RF is accurate in identifying poor households that are candidates for social assistance programs in Indonesia. Thoplan (2014) uses RF to estimate poverty in Mauritius, the results showed that RF is the most accurate method for estimating poverty. Mcbride & Nichols (2015) successfully apply RF to improve the precision of PMT compared to linear regressions. They conclude that the RF model can significantly improve

PMT performance by 2 to 18%. Another study similarly compares different data mining methods for poverty classification based on censuses and surveys applied to the population. It is concluded that RF outperforms Logistic Regression and Support Vector Machine (SVM), in addition to being relatively faster than the other methods (Korivi, 2016).

In addition to the comparison between methods to determine the most suitable for estimating poverty, there are also studies where different sources of information are used, to find patterns that allow estimating poverty more precisely. For example, Lerman et al. (2016) obtain indicators of economic income and education based on georeferenced Twitter interaction. Jean et al. (2016) use high-resolution satellite imagery in conjunction with machine learning algorithms to predict poverty in African countries. As well as Piaggesi et al. (2019) use satellite imagery data to predict urban poverty in poor countries.

Similarly, Pokhriyal (2019) uses multiple sources of information to map poverty, establishing relationships between poverty and auxiliary data sources, such as cell phone records, satellite images, weather measurements, and Open Street Maps. Poverty is approached as a regression problem where it is sought to estimate a poverty index value for each region, taking as a reference the poverty value obtained through census information.

On the other hand, Khaefi et al. (2019) identify characteristics obtained from Call Detail Records (CDR), which allow predicting wealth and poverty in New Guinea, combining CDR data with survey data. Techniques such as Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) are used to obtain the relative index of multidimensional wealth in households. In addition, it uses feature selection techniques such as Fast correlation-based filter (FCBF), Boruta, and XGBoost. Finally, five algorithms are applied to estimate the relative multidimensional wealth index. The algorithms used are SVM, NB, Elastic-net, Neural networks (NN), and Decision trees. The researchers conclude that CDRs are better suited for generating asset-related indicators and quantile rankings based on the wealth index of individual households. Howerver, they are not suitable for estimating the relative value of the wealth index.

Multidimensional Poverty

Multidimensional poverty is an approach to assess poverty that arises from the question that well-being is associated with income and consumption Añazco & Pérez (2016); under this perspective analysis, it is necessary to evaluate social welfare from new non-monetary dimensions (Denis et al., 2010). Therefore, poverty for this approach represents a situation of insufficient realization of certain capacities considered elementary (Añazco & Pérez, 2016). The multidimensional poverty index (MPI) is a methodology developed by Alkire and Foster in 2007 and from its formulation it has become the most widely used statistical practice to measure multidimensional poverty worldwide (Añazco & Pérez, 2016). This index has been incorporated into the annual Human Development reports developed by the United Nations Development Program (UNDP) from 2010 to the present. The multidimensional poverty index allows identifying the simultaneous deprivations that the individual experiences in the enjoyment of their rights (Quinde Rosales, Bucaram Leverone, Saldaña Vargas, & Martínez Murillo, 2020).

The Multidimensional Poverty Index (MPI) makes it possible to identify the simultaneous deprivations that the individual experiences in the enjoyment of their rights (Rosales et al., 2020).

Studies in Latin America regarding multidimensional poverty have been directed towards a basic set of dimensions such as education, health, employment, social protection, housing conditions, and basic services. These are consistent with the findings of participatory studies at the international, regional or national level (Clausen et al., 2019). Many countries such as Colombia (Salazar et al., 2011), Mexico (CONEVAL, 2016), Chile (Social, 2015), El Salvador (Munguía, 2017) measure the MPI based on the methodology proposed by Alkire and Foster (2011), and adopted by the Economic Commission for Latin America and the Caribbean (CEPAL) as the standard for the region.

In Ecuador, the MPI is calculated from the sociodemographic information collected through the ENEMDU household survey. The data is stored in a database that contains the answers of each respondent. Then, based on data related to education, health services, access to food and water, and social security, the MPI is calculated for each individual and then extended to the familiar unit. This survey is applied only to the main cities of the country which implies that many regions within the country and cities are not considered. Besides, this survey follows a probabilistic sampling strategy. Therefore, the overall results are just a probabilistic approximation based on the expansion factor from the sampling strategy (Añazco & Pérez, 2016).

Regarding the multidimensional poverty related to ML in Ecuador, Viscaino Caiche (2019) performs a Statistical Matching between the variables of the Living Conditions Survey (ECV) and the CPH 2010, to subsequently estimate the MPI at the regional, provincial and cantonal level. Using Decision Trees and NN, an explanatory model of poverty is found. Finally, the Poverty Index is estimated through a Logistic Regression. The results show that the province with the highest poverty rate is Morona Santiago, while Galapagos is the province with the lowest poverty rate.

Methodology

The type of study corresponds to a causal cross-sectional non-experimental quantitative design. Non-experimental designs "are carried out without the deliberate manipulation of variables and phenomena are observed in their natural environment to analyze them" (Sampieri Hernández et al, 2014, p. 152). Additionally, cross-sectional causal designs "describe relationships between two or more categories, concepts or variables at a given time based on the cause-effect relationship" (Sampieri Hernández et al, p. 158); In this sense, in the different evaluated models of machine learning, a series of predictive variables of a sociodemographic nature are included that affect multidimensional poverty.

This study starts with the collection of the databases ENEMDU and CPH 2010. Using the Political Administrative Division (DPA), which is a codification strategy that assigns an identifier for each city, area, and sector in the country. The data of Cuenca city is extracted from both databases.

Due to the CPH is carried out every 10 years, and the 2020 version is not yet available. Therefore, the 2010 version is used. In addition, the ENEMDU 2010 survey is used to coincide with the temporal context.

The databases are looked up to find common variables. Most of them are related to socio-demographic information, such as education, housing, employment, marital status, age, sex, etc. 33 common variables were found. In the process of identifying common variables between the databases, it was found that the variables related to the DPA coding of the cities, areas, and sectors were present in both databases and had the same coding. However, they were

not included as common variables to avoid the model having a bias towards the geographic location of the observation.

Each of these variables has its structure. The objective is to have the same structure between corresponding common variables, so the next step is to process them to change their original structure to match with the corresponding common variable. Thus, the ML can be used with both databases, because their input variables have the same structure. This process is described in detail in section 3.A.1.

The input features of the machine learning models must have specific characteristics to facilitate the training process. Therefore, categorical and numerical features must be preprocessed. The method is described in detail in section 3.A.2. After this procedure, 140 input features were obtained.

Once the data is ready, the 3440 available observations from the ENEMDU database, are split into training and test datasets, with a ratio of 80% training and 20% test, following a random sampling strategy.

Ten ML models, from simple linear regressors to ensemble models like RF, were trained to find the most suitable for estimating the MPI using the common variables as input features. This process is explained in detail in section 3.B.

The models with good performance are further tested adjusting their hyperparameters using a GridSearch strategy, which tries to find the best combination of parameter's values from given distributions, testing every single combination. This is achieved by first picking a combination of hyper-parameters. Then, with those hyperparameters, a training and cross-validation process is performed. This is done for each combination. After all the combinations are trained and validated, the one with the best cross-validation performance is chosen as the best model (scikit-learn, 2021).

Finally, the best model can be used to estimate the MPI using the CPH common variables as input features. Thus, we have the MPI for every observation in the CPH database, without any region excluded.

Data preprocessing
Recoding
To have the same structure between the common variables on the ENEMDU and CPH database, many variables must be recoded. Table 1 shows an example for a variable that represents the educational level. It can be noticed that option 3 from both databases, which represents the kinder garden, is merged with option 4 that represents primary education. Option 8: "Ciclo Postbachillerato" from the ENEMDU database is merged with option 7: "Superior no Universitario" of the new coding proposal. Also, the order in the options for both databases is changed to match the proposed recodification.

The same strategy is applied to all the variables that have a different structure from their correspondent common counterpart.

It is important to mentions that during this process it was found that some of the variables contained an assigned category (number 99) for when the respondent cannot answer the question. One might think that these values do not represent relevant information and should be discarded from the analysis. However, the fact that the respondent does not answer

may have deeper connotations. For example, if the respondent is not able to answer "how many rooms are destined only for resting?" it may imply an overcrowding situation. That is why null values must also be taken into account as a source of information. Therefore, when recoding the variables, there must be a coding for these values.

Table 1

Recoding Variable "Level of Instruction"

ENEMDU		CPH		
Variable	*Options*	*Variable*	*Options*	*New coding*
What is the highest level of instruction you attend or attended?	1.0: 'None', 2.0: 'Literacy Center', 3.0: 'Preschool', 4.0: 'Primary school', 5.0: 'Highschool', 6.0: 'Basic education', 7.0: 'Middle education', 8.0: 'Post-Baccalaureate Cycle', 9.0: 'Higher education', 10.0: 'Postgraduate', 99.0: 'It is ignored'	Level of instruction	1.0: 'None', 2.0: 'Literacy Center', 3.0: 'kinder garden', 4.0: 'Primary school', 5.0: 'Basic education', 6.0: 'Highschool', 7.0: 'Middle education', 8.0: 'Non-university superior', 9.0: 'University Superior', 10.0: 'Postgraduate'	1.0: 'None', 2.0: 'Literacy Center', 3.0: 'Primary school', 4.0: 'Basic education', 5.0: 'Highschool', 6.0: 'Middle education', 7.0: 'Non-university superior', 8.0: 'University Superior', 9.0: 'Postgraduate'

Source: The authors
Sources: ENEMDU(2010), CPH (2010)

Data Conditioning

This study considers two kinds of variables: numerical variables and categorical variables. Each of them needs their conditioning to simplify its representations and help in the training process because most training algorithms are optimized to use only numbers as input features, they could binary or decimal points (Géron, 2019).

For numerical features such as the age or the number of bedrooms in a house, the conditioning is called feature scaling. This is done because most ML models do not perform well when their input has different scales. There are two common methods to scale the features. 1) mix-max scaling and 2) standardization.

mix-max scaling (a.k.a. normalization) consists of shifting and rescaling the values so that they range from 0 to 1. This is achieved by subtracting the minimum value and dividing the maximum minus the minimum. Standardization on the other hand first subtracts the mean (to have a zero-mean distribution) and then divides the result by its standard deviation (to have unit variance) (Géron, 2019).

In this study, the standardization is used through the StandardScaler function of the sklearn library.

For categorical features, many strategies can be applied. The most popular are Ordinal Encoding and One Hot Encoding. Both strategies convert a categorical variable into numbers. Ordinal Encoding assigns a number for each category. This can be used with ordered categories such as bad, average, good, excellent. Not all the categorical variables represent an ordered list. In that case, the One Hot Encoding must be used. This strategy creates one binary variable for each category, being 1 only the binary variable that corresponds with the category that represents. Table 2 shows an example of the One Hot Encoding strategy.

Table 2

One hot Encoding

Categorical Variable	One Hot Encoding			
Values	Cat A	Cat B	Cat C	Cat D
A	1	0	0	0
D	0	0	0	1
C	0	0	1	0
B	0	1	0	0

Source: The authors.

It can be noticed that one categorical variable with four categories is transformed into 4 binary variables. For all the categorical variables in our study, the One Hot Encoding strategy is used through the OneHotEncoder function of the sklearn library. After the feature conditioning process, we pass from 33 to 140 variables

Model Exploration

To evaluate each model, a performance criterion must be chosen. This criterion usually represents the distance between two vectors (a.k.a. *norm*), the vector of prediction, and the target vector. For regression models, the most common criterion is the *Euclidean distance*, which is calculated by the Root Mean Square Error (RMSE), also called the l_2 norm. RMSE assigns a higher weight for larger errors. Another distance measure is the *Manhattan norm* calculated by the Mean Absolute Error (MAE), also called the l_1 norm. The higher the norm index, the more focus is on large errors and neglects small ones. That is why the RMSE is more sensitive to outliers than the MAE, but when there are just a few outliers RMSE is preferred (Géron, 2019).

For this study RMSE is used because the aim is to penalize large prediction errors, the outliers are quite rare, and because it uses the same units as the dependent variable (MPI). So, the results are easy to compare.

Below, the process for training a Linear Regressor is shown. The same steps are followed for all the other models.

1. Train the linear regressor with default hyperparameters from the python library sklearn as shown in Table 3.

Table 3

Linear Regressor Hyper-Parameters

| Hyper Parameter	Value
fit intercept	True
normalize	False
positive	False

Source: The authors

2. Calculate the RMSE. In theory, the MPI can have values between 0 and 1. Under this condition, an RMSE of 0.0855, as shown in Table 4, is equal to a relative error of 8.55%. However, the real range of the MPI in the dataset is between 0 and 0.7917. So, the relative error is 10.80%. The second approach is better because considers the real data, and avoid overestimating the performance of the model.

Table 4

Training Performance Measures of the Linear Regressor

RMSE	Range	% error
0.0855	0.7917	10.80%

Source: The authors

3. Generate a scatter plot to visualize the predictions on the training set as shown in Figure 1. As the dataset is sorted according to the survey sampling strategy it would be very difficult to visualize. For that reason, the samples are sorted in ascending order. The target MPI from the training dataset is plotted in green and the MPI estimated by the linear regression is plotted in red.

Figure 1

Estimated MPI vs Target MPI with the Linear Regressor from the Training Dataset
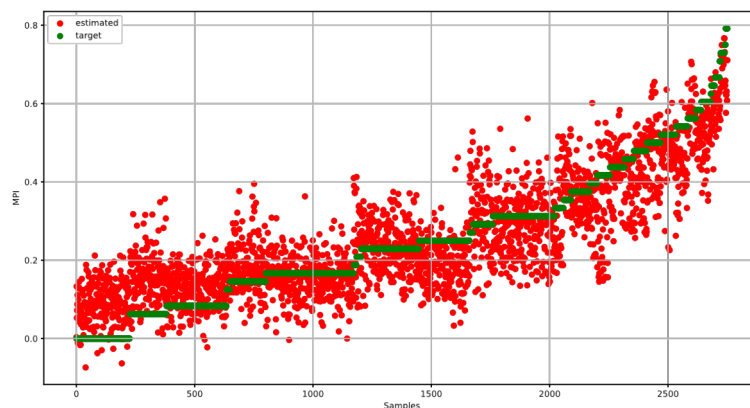
Source: The authors

4. Perform cross-validation with 10 folds. The cross-validation method with 10 folds equally divides the dataset into 10 small datasets following a random sampling strategy. Then, the first nine datasets are used to train the model, while the 10th remaining dataset is used to test the performance. The process is repeated until each dataset has been used for testing. The RMSE for each fold is calculated. The mean of all the RMSE is calculated to obtain an overall

performance measurement that considers each iteration of the cross-validation process. Also, the relative error is calculated from the range of the MPI in the dataset.

Table 5

Cross-validation Performance Measures of the Linear Regressor

Mean RMSE	Range	% error
0.0905	0.7917	11.43%

Source: The authors

To be able to compare the models, it is necessary to define the concepts of underfitting and overfitting. This allows understanding the overall performance of the model and permits the proposal of some solutions to improve it.

Comparison between models

When a model is underfitting means that is not capable to make good estimations, mainly for two reasons. 1) The features do not provide enough information or 2) the model is not powerful enough. To fix the underfitting problem we can select a more powerful model, feed the training algorithm with better features, or reduce the constraints of the model (Géron, 2019). On the other hand, when a model is overfitting means that the model is capable to make good estimations, but only with the same data with which it was trained. When trying to estimate with different data the performance is bad. To determine if the model is overfitting it is necessary to compare the performance with training data vs the performance with unseen data. As the unseen data is not available, a way to estimate its performance is through a cross-validation process. If the training performance is much better than the cross-validation performance, the model is overfitting. Overfitting happens when the model is too complex to the amount or the noisiness of the data. To fix this problem we can simplify the models, constrain them (also called regularization), reduce the noise in the data, or feed more training data (Géron, 2019).

At this stage, multiple ML models are evaluated and compared by overfitting or underfitting analysis to identify the model that best performs in the task of estimating the MPI. The tested models are:

- Linear Regressor
- Polynomial Regressor
- Ridge Regressor
- Bayesian Regressor
- Elastic-net Regressor
- Linear SVM
- SVM, kernel: Polinomial (d=3)
- SVM, Kernel: RBF
- Decision Tree
- Random Forest

The performance of each of these models is evaluated in section 4.

Fine Tuning

Fine Tuning is the process where the hyper-parameters are adjusted to obtain the best possible model (Géron, 2019). First, it is necessary to determine the hyper-parameters available to tweak from the model.

For this case, Random Forest was chosen as the best model. So, its hyper-parameters are (scikit-learn, 2021):

max depth: The maximum depth of the tree.
max features: The number of features to consider when looking for the best split.
bootstrap: The sampling strategy. If False, the whole dataset is used to build each tree. If true, a bootstrap strategy is applied.

There are many other hyper-parameters, but those were not considered in this study.
With these hyper-parameters, an exploration grid is created. This grid contains all the possible combinations that can be explored by the search algorithm.

Figure 2

Estimated MPI vs Target MPI with the Random Forest Regressor from the Training Dataset
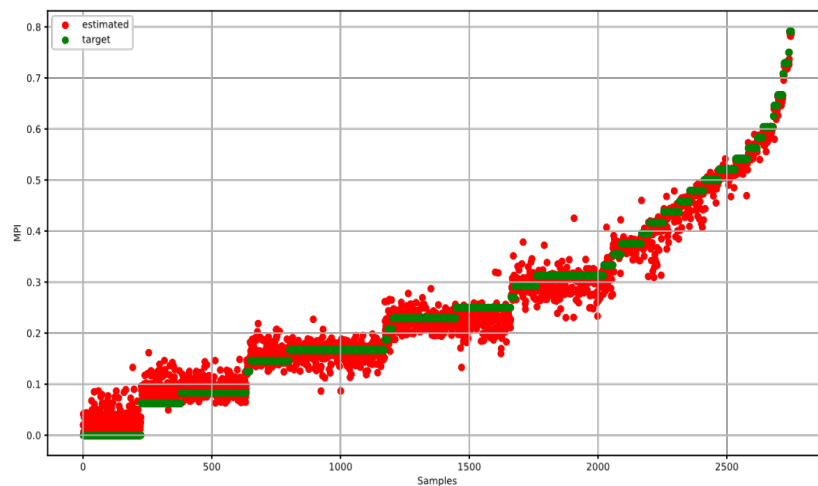
Source: The authors

As mentioned in section 3, the GridSearch strategy is used to explore the possible combinations of hyper-parameters. The exploration grid is shown in Table 6.

Table 6

Hyper-parameters Exploration Grid

Hyper-parameter	Values
n estimators	10, 50, 100, 500, 1000
max depth	5, 10, 50, 100, 500, 1000
max features	10, 50, 100, 140
bootstrap	True, False

Source: The authors

According to the number of values for each hyper-parameter in the exploration grid, there are 240 possible combinations. All of them are tested by the GridSearch algorithm. In addition, with each combination, 10-fold cross-validation is performed. So, the training process is executed 2400 times. The combination with the best performance is chosen as the best model. The best combination of hyperparameters is shown in Table 7.

Table 7

Best Combination of Hyper-Parameters

n estimators	max features	max depth	bootstrap
1000	50	500	False

Source: The authors

The performance is measured through the mean RMSE of the 10-fold cross-validation and the percentage of error based on the MPI range, as shown in Table 8.

Table 8

Performance Measures of the best Random Forest Model

Mean RMSE	Range	% error
0.0594	0.7917	7.50%

Source: The authors

It can be noticed that after adjusting the hyper-parameters the cross-validation performance of the model was improved, going from an error of 8.95% to 7.50%.

Results

Table 9 shows the results obtained from all the tested models. It can be noticed that Linear Regressor has a consistent performance along with training and cross-validation. However, the error is around 11%, which is considered under-fitted. Polynomial Regressor, is overfitted because the cross-validation error is much higher than the training error. This can be solved using a regularized model such as Ridge Regressor, Bayesian Regressor, or Elastic-Net. The Ridge Regressor shows better performance on the cross-validation but is still overfitted. Bayesian Regressor has a good performance overall and shows consistency between training and cross-validation performance. Elastic-net shows consistency between training and cross-validation but is still under-fitted with an error of 11%.

Continuing with other models, Linear SVM is under-fitted with an error of around 20%. For no Linear SVM, two kernels were tested. 1) Polynomial Kernel with a degree of 3 and 2) Radial basis function (RBF) Kernel. SVM with Polynomial kernel shows good performance but it is still considered under-fitted with a cross-validation error of 10%. Similarly, SVM with RBF kernel has an error of 10% on the cross-validation. On the other hand, the Decision Tree is also highly overfitted with 0.5% of training error and 11% of the cross-validation error. Finally, Random Forest shows the best cross-validations performance among all of them. It is quite overfitted with a training error of around 3% and a cross-validation error of 8.95% but this can be solved by adjusting the hyper-parameters. Thus, Random Forest is chosen as the best model, Figure 2 shown the plot from this model.

Table 9

Comparison of Models Performance with Training data set

Model	Training		Cross-Validation	
	RMSE	% error	Mean RMSE	% error
Linear Regressor	0.0911	11.5	0.0904	11.42
Polynomial Regressor	0.0117	1.48	1.6552	209.07
Ridge Regressor	0.0135	1.70	0.0758	9.57
Bayesian Regressor	0.0518	6.55	0.0742	9.36
Elastic-net Regressor	0.0901	11.38	0.0923	11.65
Linear SVM	0.1651	20.86	0.1653	20.88
SVM, kernel: Polinomial (d=3)	0.0683	8.63	0.0807	10.19
SVM, Kernel: RBF	0.0693	8.75	0.0809	10.22
Decision Tree	0.0046	0.58	0.0874	11.04
Random Forest	0.0288	3.64	0.0709	8.95

Source: The authors

As mentioned above, the main purpose of this study is to estimate the MPI through an ML model using sociodemographic variables as inputs. With the trained model, now we can estimate the MPI for new unseen data. We expect an error similar to the one obtained with the cross-validation process.

Usually, the only way to evaluate the final performance of the model is through the reserved test dataset. This is described in section 4.A. However, in this particular case, it is available another way to evaluate the performance based on the comparison between the overall MPI indexes of the ENEMDU database as the ground truth and the overall MPI indexes calculated from the estimated MPI of the CPH database. This is described in detail in section 4.B.

Test Dataset

Using the reserved 20% of the data to evaluate the final performance of the model has the advantage of having the target value of MPI to be compared with. So, we can plot the estimated MPI vs the target MPI, as shown in Figure 3.

The measures of RMSE and the percentage of error, between the estimated MPI and the target MPI from the test dataset, are shown in Table 10.

Table 10

Performance Measures with the test Dataset

| Mean RMSE	Range	% error
0.0592	0.7917	7.48%

Source: The authors

Overall MPI

When the model is used with the CPH database, we do not have the expected value to compare with. Therefore, the RMSE and the percentage of error cannot be calculated.

However, the performance can be approximately measured if the overall results between the estimated MPI from the CPH database and the target MPI from the ENEMDU database are compared. 4 indexes that represent a summary of the multidimensional poverty in a region can be calculated. These 4 indexes are:

- MPR: Multidimensional Poverty Ratio. The percentage of the population with Multidimensional Poverty. *MPI* >= 0.33.
- EMPR: Extreme Multidimensional Poverty Ratio. The percentage of the population with Extreme Multidimensional Poverty. *MPI* >= 0.5.
- A: The average MPI among the population with Multidimensional Poverty.
- MPI: The average MPI in the region.

Figure 1

Estimated MPI vs Target MPI with the test Dataset
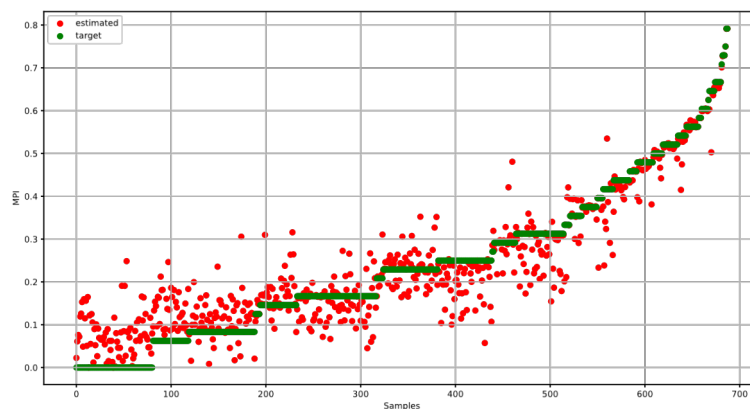
Source: The authors

For Cuenca city, Table 11 shows a comparison between the target indexes from the ENEMDU database vs the estimated indexes from the CPH database using the tuned RF model.

Table 11

Target vs Estimated Multidimensional Poverty in Cuenca

	ENEMDU		CPH		
	Negative	**Positive**	**Negative**	**Positive**	**Difference**
MPR	74.53%	25.47%	74.25%	25.75%	0.28%
EMPR	88.41%	11.59%	95.15%	4.85%	6.74%
A	0.4877		0.4245		0.063
MPI	0.1242		0.1092		0.015

Source: The authors

It can be noticed, that the estimated indexes are very similar to the target indexes. The difference in MPR is only 0.28%, which means that the model estimates almost the same value of MPI close to the multidimensional poverty threshold (0.33). The difference in EMPR (6.74%) shows that the model estimated considerably higher values of MPI close to the extreme multidimensional poverty threshold (0.5). About the average MPI among the

population with multidimensional poverty (A), the difference is only 0.0028, which represents 7.95% of the MPI range. Overall MPI in the region differs by 0.015, which is 1.89% of the range.

Discussion

All the analysis and results of this study move around the data, so its preprocessing is essential for the study. This preprocessing not only covers technical aspects such as the cleaning of the database, the imputation of missing values, or the conditioning of variables but also has a strong component of theoretical concepts that must be taken into account to maintain or discard important features.

Talking about those theoretical concepts, it is identified that the possibility of measuring the MPI through the ML, opens the range of possibilities to continue rethinking the established concepts according to social dynamics and its contemporary problems, integrating multiple variables of analysis that make visible no only those aspects that have traditionally been measured, but others that determine "poverty" in different contexts and are from a more subjective nature, such as the level of access to justice, the level of social cohesion of a territory, or the time devoted to activities considered a priority for people.

When analyzing the target MPI values (plotted in green) in Figure 1 it can be seen that the statistical calculation that yields these values works through thresholds. That is why when ordering the data ascendingly, steps defined with a constant value of MPI can be observed. This calculation methodology hides the random component of the observations. On the other hand, however, the features used as inputs for the training process maintain the typical random nature of a survey. For this reason, it is quite difficult for the models to generate estimates close to the target values. This may be one of the reasons why RF performs better than other models. Since RF, internally uses decision trees and they work precisely with thresholds for regression tasks.

Another model that performed well is the Bayesian Regressor. However, when adjusting its hyper-parameters, the improvement was not significant. That is why a deep fine-tuning analysis has not been included.

On the other hand, it can be noted that the models with the worst performance are those of a linear nature. This is an important observation since most sociological studies start with linear models to explain the phenomena, which would be an error in the case of multidimensional poverty.

The performance obtained by RF, both in the cross-validation (7.5%) and in the test dataset (7.48%) is considered acceptable given that, from a theoretical point of view, multidimensional poverty is a complex phenomenon that can have several edges and variables that have not been considered in this study.

Finally, it is necessary to analyze the comparison of the indices calculated based on the estimates made by the model with the CPH data against the indices calculated statistically from the ENEMDU database. Although the indices of the ENEMDU cannot be considered a ground truth, since they are calculated by applying the expansion factor of the survey, they can give us an approximation about the performance of our model. If the indices calculated based on our estimates are similar to those estimated statistically, it can be said that the model is fulfilling its purpose of estimating, as precisely as possible, the MPI for the entire population of a specific region. When observing the results, it can be seen that this is the case, the difference of the

average value of MPI in the entire region is 0.015, a difference of only 1.86 % of the range. However, it has to be noted that the model is very accurate close to the Multidimensional Poverty threshold (0.33) but its accurate decrease is close to the Extreme Multidimensional Poverty threshold (0.5).

Currently, there are just a few studies that address the estimation of multidimensional poverty using machine learning techniques. The study that most closely resembles ours is the one carried out by Viscaino Caiche (2019), where it performs a statistical matching between the variables of the Living Conditions Survey (ECV) and the CPH 2010. Unlike our study that uses regression models to estimate a numerical value of IPM, Viscaino Caiche (2019) uses decision trees, neural networks, and logistic regression to obtain a classification model between rich and poor. Obtaining a precision of 80.84% for the decision tree model, 83.65% for the artificial neural network model, and 83.55% for the logistic regression model. On the other hand, our regression model has an approximate accuracy of 93%.

Conclusions

This study has shown that through ML techniques, substantial aspects such as the territorialization of MPI over a meso and micro-scale is a reality, which contributes decisively to the development of science in the area of human well-being and the decision making in public policy.

In Ecuador, multidimensional poverty has been estimated only in large territorial areas and has not reached the canton level due to the lack of precise information on aspects inherent to said estimation. The application of the Random Forest model that has been tested in this study has given the possibility of estimating the MPI of individuals at the sector level in the Cuenca canton, which implies a substantial advance for the visibility of poverty in its multiple dimensions as one of the most complex social problems in the country. In addition, the model has managed to cover the MPI both in the urban area and in the rural area, which will help to analyze the impact of the differences underlying the provision of services and the guarantee of rights in the territories as a substantive input for making decisions of the decentralized autonomous governments.

This research finding is also relevant insofar as it manages to establish multidimensional poverty levels and identify extreme multidimensional poverty, with which the target populations can be taken into account to focus affirmative actions and urgent interventions on the less favored territories. as well as long-range policies that make it possible to achieve a model of sustained territorial equity.

The robustness of the Random Forest model and its application in the estimation of MPI proven through the validation of accuracy with other similar models will allow it to be applied in different latitudes with the same success, which is why it constitutes a basic tool for academia and public management.

In the information age, the social sciences not only have the opportunity, but also the obligation to adopt new technologies in their research. It is well known that interdisciplinary work has led to great advances as a society. It is time for the social sciences to work together with other sciences. Not only between sciences of the same family, such as statistics and economics but also with sciences such as engineering to bring technical advances to the social field and in the same way social advances to technical fields. In this way, both fields benefit. The social sciences become more precise in their theories and engineering becomes more humane.

After this first approach between data analysis techniques and the social sciences, it can be noted that joint work allows carrying out researches that would be very difficult separately. As seen in the results, the ML has great potential as support in studies of a social nature. Mitigating one of the main problems of the social sciences such as the scarcity of data.

It is important to mention that ML algorithms are only tools, which by themselves will not solve the problems of the social sciences. ML will perform well as long as there is strong theoretical support and it is applied to appropriate research problems. The theory of social sciences should be the guide for the research design and the interpretation of the results. A guide that the data alone will never provide.

The tasks of data cleaning, model training, and prediction of the MPI can be automated since they are nothing more than programming codes that are executed sequentially. However, there is a large part of this work that cannot be automated since it requires the exclusive intervention of a person, specifically in the task of identifying and recoding common variables. These tasks require a sense of association between the information provided by each variable and knowledge of their meaning, tasks that for now cannot be executed by a computer.

There is still a long way to go in this bridge formed between these sciences, the scarcity of data is not the only problem to be solved and multidimensional poverty is not the only variable of interest. It is still difficult to analyze all the information found in this infinite source of information such as social networks. Phenomena such as xenophobia, racism, misogyny, violence, security, public policies, among others, are issues that can be addressed from data science in conjunction with social sciences. This cooperative work must be maintained. Therefore, universities, companies, and society, in general, must promote this type of multidisciplinary study.

References
Alkire, S., & Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of Public Economics*, *95*(7-8), 476–487. https://doi.org/ 10.1016/j.jpubeco.2010.11.006

Añazco, R. C., & Pérez, F. J. (2016). Medición de la Pobreza Multidimensional en Ecuador. *Revista de Estadística y Metodología*, 27–51.

Chen, N. C., Drouhard, M., Kocielnik, R., Suh, J., & Aragon, C. R. (2018). Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems*, *8*(2). https://doi.org/10.1145/3185515

Clausen, J., Vargas, S., & Barrantes, N. (2019). Do official multidimensional poverty measures in Latin America reflect the priorities of people living in poverty? *Ensayos de Política Económica*, *2*(6), 15–34.

Consejo Nacional de Evaluación de la Política de Desarrollo Social. (2016). *Metodología para la medición multidimensional de la pobreza en México*. https://www.coneval.org.mx/Medicion/MP/Paginas/Metodologia.aspx

Denis, A., Gallegos, F., & Sanhueza, C. (2010). Medición de pobreza multidimensional en Chile. Santiago de Chile: Universidad Alberto Hurtado.

Devarajan, S. (2013). Africa's Statistical Tragedy. *Review of Income and Wealth*, *59*(SUPPL1), S9–S15. https://doi.org/10.1111/roiw.12013

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS, Political Science & Politics*, *48*(1), 80.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science, 24*(1), 395–419. https://doi.org/10.1146/annurev-polisci-053119015921

Hand, D. J. (2007). Principles of data mining. *Drug Safety, 30*(7), 621–622. https://doi.org/10.2165/00002018-200730070-00010

Hindman, M. (2015). Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. *Annals of the American Academy of Political and Social Science*, 659(1), 48–62. https://doi.org/10.1177/ 0002716215570279

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790–794. https://doi.org/10.1126/science.aaf7894

Kambuya, P. (2020). Better Model Selection for Poverty Targeting through Machine Learning: A Case Study in Thailand. T*hailand and The World Economy, 38*(1), 91–116.

Khaefi, M. R., Hendrik, Burra, D. D., Dianco, R. F., Alkarisya, D. M. P., Muztahid, M. R., Zahara, A., Hodge, G., & Idzalika, R. (2019). Modelling Wealth from Call Detail Records and Survey Data with Machine Learning: Evidence from Papua New Guinea. Proceedings - 2019 *IEEE International Conference on Big Data, Big Data 2019*, 2855–2864. https: //doi.org/10.1109/BigData47090.2019.9005519

King, G., Keohane, R. O., & Verba, S. (1995). The importance of research design in political science. *American Political Science Review*, 89(2), 475-481.

Korivi, K. (2016). *Identifying poverty-driven need by augmenting census and community survey data*. [master's thesis, Kansas State University]. Institutional Repository UN. http://hdl.handle.net/2097/34556

Kshirsagar, V., Wieczorek, J., Ramanathan, S., & Wells, R. (2017). *Household poverty classification in data-scarce environments: a machine learning approach*. arXiv. http://arxiv.org/abs/1711.06813

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., & Gutmann, M. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915), 721.

Lerman, K., Arora, M., Gallegos, L., Kumaraguru, P., & Garcia, D. (2016). *Emotions, Demographics and Sociability in Twitter Interactions* (tech. rep. No. 1). http://sentistrength.wlv.ac.uk/

Maldonado, C. E. (2019). Three reasons for social sciences metamorphosis in the 21st century. *Cinta de Moebio*, *64*(64), 114–122. https://doi.org/10. 4067/S0717-554X2019000100114

Mcbride, L., & Nichols, A. (2015). *Improved poverty targeting through machine learning: An application to the USAID Poverty Assessment Tools* (tech. rep.).

Molina, M., & Garip, F. (2019). Machine Learning for Sociology. *Annual Review of Sociology*, *45*, 27–45. https://doi.org/10.1146/annurev-soc-073117041106

Moreno, M. (2017). La medición de la pobreza. *Revista Sociedad*, (37).

Munguía, F. (2017). Medición multidimensional de la pobreza: El Salvador. In Villatoro, P. (Comp.), *Indicadores no monetarios de pobreza: avances y desafíos para su medición. (*págs.105-109). Comisión Económica para América Latina y El Caribe.

Otok, B. W., & Seftiana, D. (2014). *The Classification of Poor Households in Jombang With Random Forest Classification And Regression Trees (RF-CART) Approach as the Solution In Achieving the 2015 Indonesian MDGs' Targets* (tech. rep.). www.ijsr.net

Piaggesi, S., Gauvin, L., Tizzoni, M., Adler, N., Verhulst, S., Young, A., Price, R., Ferres, L., Cattuto, C., & Panisson, A. (2019). *Predicting City Poverty Using Satellite Imagery* (tech. rep.). https://censusreporter.org/topics/ income/

Pokhriyal, N. (2019). Multi-View learning from disparate sources for poverty mapping. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, *33*(01), 9892–9893. https://doi.org/10.1609/aaai.v33i01.33019892

Rosales, V. Q., Leverone, M. B., Vargas, M. S., & Murillo, C. M. (2020). Multidimensional Poverty Index and its relationship with Ecuadorian public spending. *Universidad y Sociedad*, *12*(2), 430–436.

Rovai, A. P., Baker, J. D., & Ponton, M. K. (2013). *Social science research design and statistics: A practitioner's guide to research methods and IBM SPSS*. Watertree Press LLC.

Rudin, C. (February, 21, 2015). Can Machine Learning Be Useful for Social Science? http://citiespapers.ssrc.org/can-machine-learning-be-useful-for-socialscience/

Salazar, A., Cuervo, Y. D., & Pinzón, R. P. (2011). ´Índice de pobreza multidimensional para Colombia (IPM-Colombia) 1997-2010. *Archivos de economía*, 382.

Sampieri Hernández, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación*. México D.F.: Mc Graw Hill.

scikit-learn. (2021). sklearn.ensemble.RandomForestRegressor. https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

scikit-learn. (2021). sklearn.model selection.GridSearchCV. https://scikit-learn.org/stable/modules/generated/sklearn.model%5C selection.GridSearchCV. html

Social, O. (2015). Nueva metodología de medición de la pobreza por ingresos y multidimensional. Ministerio de Desarrollo Social.

Sohnesen, T. P., & Stender, N. (2017). Is ROom Forest a Superior Methodology for Predicting Poverty? An Empirical Assessment. *Poverty & Public Policy*, *9*(1), 118–133. https://doi.org/10.1002/pop4.169

Talingdan, J. A. (2019). Performance comparison of different classification algorithms for household poverty classification. *Proceedings - 2019 4th International Conference on Information Systems Engineering, ICISE 2019*, 11–15. https://doi.org/10.1109/ICISE.2019.00010

Thoplan, R. (2014). Random Forests for Poverty Classification. International Journal of Sciences: Basic and Applied Research. 17(2).

Viscaino Caiche, L. (2019). *Estimación de Índice de pobreza multidimensional a nivel provincial para Ecuador.[master's thesis, Universidad de Cantabria, España]. Institutional Repository UN.* http://hdl.handle.net/10902/18129

Wallach, H. (2016). *Computational social science: discovery and prediction. Cambridge University Press.*