Estudios Económicos (México, D.F.)
ISSN: 0188-6916
El Colegio de México A.C.

Rincón, Ratzanyel
Quarterly multidimensional poverty estimates in Mexico using machine learning algorithms
Estudios Económicos (México, D.F.), vol. 38, no. 1, 2023, January-June, pp. 3-68
El Colegio de México A.C.

DOI: https://doi.org/10.24201/ee.v38i1.435

QUARTERLY MULTIDIMENSIONAL POVERTY ESTIMATES IN MEXICO USING MACHINE LEARNING ALGORITHMS

ESTIMACIONES TRIMESTRALES DE POBREZA MULTIDIMENSIONAL EN MÉXICO MEDIANTE ALGORITMOS DE APRENDIZAJE DE MÁQUINA

Ratzanyel Rincón

The University of British Columbia

Resumen: Este artículo aborda la falta de información oportuna sobre la pobreza multidimensional en México. Tres algoritmos de aprendizaje de máquinala regresión LASSO logística, el bosque aleatorio y las máquinas de vectores de soporteson entrenados con la ENIGH para encontrar patrones generalizables de pobreza multidimensional en los datos. Los modelos se utilizan para clasificar a cada individuo en la ENOE como pobre o no-pobre para obtener tasas de pobreza trimestrales. Estas estimaciones son más cercanas a los niveles de pobreza multidimensional que la pobreza laboral y brindan una perspectiva precisa sobre la pobreza con más de un año de antelación a la medición oficial.

Abstract: This article addresses the lack of timely information about multidimensional poverty in Mexico. Three machine learning algorithmsthe LASSO logistic regression, random forest, and support vector machinesare trained with the ENIGH to find generalizable patterns of multidimensional poverty in the raw data. The fitted models are used to classify each individual in the ENOE as poor or non-poor to obtain aggregated poverty rates on a quarterly basis. These estimates are closer to the official levels of multidimensional poverty than the labor poverty measurement and provide an accurate poverty outlook more than a year ahead of the official measure.

Clasificación JEL/JEL Classification: C140, C8, D6, I320

Palabras clave/keywords: multidimensional poverty; machine learning; LASSO logistic regression; random forest; support vector machines

Fecha de recepción: 14 VII 2021 *Fecha de aceptación: 30 VIII 2021*

https://doi.org/10.24201/ee.v38i1.435

1. Introduction

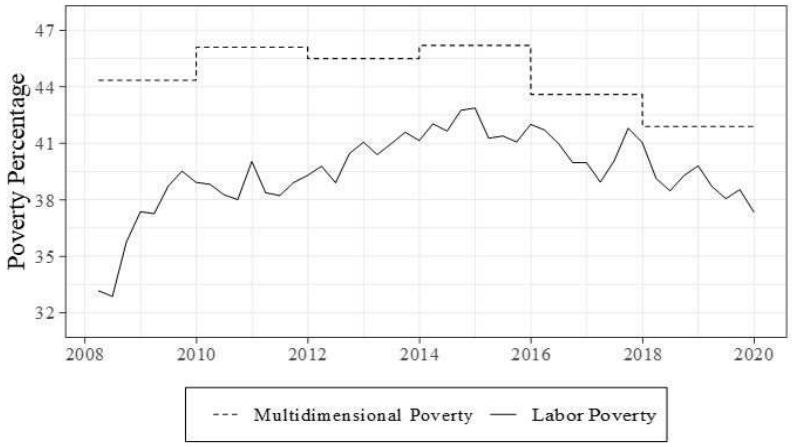
Measuring poverty is crucial to developing countries and to international organizations such as the World Bank, the OECD, and the IDB. These organizations have emphasized that the proper measurement of poverty is as important as fighting it, mainly because addressing poverty is not only about designing public policy, but also about accurately assessing its performance to examine its determinants, make budgetary decisions, and study its behavior during economic crises. Poverty measurement helps developing countries to gauge program effectiveness and guide development strategy in a rapidly changing economic environment (World Bank, 2015).

Mexico lacks high-frequency information about multidimensional poverty because of the biennial periodicity of the official measurement, which has been provided since 2008 by the National Council for the Evaluation of Social Development Policy (*Consejo Nacional de Evaluacion de la Politica de Desarrollo Social*, CONEVAL), based on the National Survey of Household Income and Expenditure (*Encuesta Nacional de Ingresos y Gastos de los Hogares*, ENIGH) and its Socioeconomic Conditions Module (*Módulo de Condiciones Socioeconómicas*, MCS).[1] CONEVAL also publishes a quarterly rate of labor poverty, estimated for the Index of Labor Poverty Trends (*Índice de la Tendencia Laboral de la Pobreza*, ITLP), which is based on the National Occupation and Employment Survey (*Encuesta Nacional de Ocupación y Empleo*, ENOE). However, CONEVAL clarifies that this estimation is not an official measure of poverty, and it does not reliably estimate multidimensional poverty on a quarterly basis. In particular, CONEVAL (2010) specifies that a person is poor according to the ITLP if their monthly per capita labor income is below the quarterly average of the minimum welfare line (mwl), whereas multidimensional poverty is based on the welfare line (which is greater than the mwl) and per capita total income (which is greater than or equal to per capita labor income). The measures cannot be directly compared, since a person identified as multidimensionally poor might have a labor income just above the mwl, but a person in labor poverty may

[1] Following enactment of Article 36 of the General Law of Social Development (Cámara de Diputados, 2004), CONEVAL developed the official Methodology for Multidimensional Poverty Measurement in Mexico (CONEVAL, 2014), which defines a person as multidimensionally poor if their total income is below the welfare line and they experience at least one of the following six social deficiencies: educational lag, lack of access to health care , lack of access to social security, inadequate housing, lack of basic household services, or inadequate nutrition.

not have a deficiency in other social dimensions or might have other income sources that boost their total income above the welfare line.[2] Figure 1 shows how labor poverty based on the ITLP has permanently underestimated the multidimensional poverty rate since 2008.

Figure 1
National labor poverty and official multidimensional poverty



Source: CONEVAL poverty estimates from the first quarter of 2008 to the last quarter of 2019 (CONEVAL, 2019, 2020).

This paper addresses the question of whether innovative, efficient, and low-cost statistical models could be applied to develop a high-frequency poverty measurement that correctly reflects the dynamics of the official multidimensional measurement in Mexico. It takes advantage of recent advancements in machine learning (ML) techniques to explore the performance and predictive power of three different classification algorithms: the LASSO logistic regression (LASSO), the random forest (RF) method, and support vector machines (SVM).[3]

[2] The minimum welfare line is equivalent to the real total value of the basic food basket per person per month, while the welfare line is equal to that amount plus the basic non-food basket.

[3] There is an extensive literature assessing predictive improvements in ML techniques in various areas of research. For instance, in health care, Cruz and Wishart (2006) find that ML algorithms increase accuracy in predicting susceptibility to cancer from 15% to 25%, in addition to providing a better understanding

The empirical strategy lies in identifying similar variables for multidimensional poverty detection in the MCS-ENIGH and the ENOE, and training LASSO, RF, and SVM to find generalizable poverty patterns in the raw data. These fitted models are then used to classify each individual in the ENOE as poor or non-poor to obtain an aggregated national poverty rate on a quarterly basis.

ML models generated with the 2008 MCS-ENIGH (training data) are used to estimate multidimensional poverty in the ENOE (prediction data) from the first quarter of 2008 through the last quarter available before publication of the 2010 multidimensional poverty measure. Similarly, models fitted with 2010 training data are used to predict the poverty in ENOE observations from the first quarter of 2010 to the last available quarter before publication of the 2012 multidimensional poverty rate, and so on. With this approach, each algorithm trained in year provides ex-post and ex-ante estimates depending on their publication date. The former corresponds to quarterly estimates during , i.e., those within the same biennial window as the training data. The latter are those from through the next CONEVAL announcement, which can be seen as predictions for the subsequent multidimensional poverty measure. Finally, these ML estimates are compared with a logistic regression (logit) model to assess their performance relative to a traditional approach.

The main results with the ML algorithms show estimates of poverty that are closer to the multidimensional poverty rate than labor poverty based on the ITLP, and they reveal dynamics that would not be seen using current measures. For instance, using ML ex-ante estimates to explore poverty dynamics during the COVID-19 pandemic, there is an increase from 3.8 to 6.4 pp in the third quarter of 2020, almost twice as much as the estimated jump during the 2008-2009 world financial crisis. However, in the fourth quarter, there is a decrease of 2.7 pp, revealing a speedy partial recovery in late 2020. These ex-ante estimates give an accurate poverty outlook more than a year ahead of the CONEVAL biennial poverty measure. The RF ex-ante and ex-post estimates are the most consistent overall, with an average gap of 0.5

of cancer development and progression. Guo *et al.* (2001) show that SVM had lower error rates in two facial recognition experiments (3.0% and 8.8%, respectively) than standard techniques. Kleinberg *et al.* (2018) describe large potential advantages in the New York City criminal justice system through use of an ML algorithmic rule, which could reduce crime up to 24.7% at the same rate of imprisonment imposed by judges, or reduce imprisonment rates up to 41.9%, without any increase in crime.

pp. Comparing out-of-sample performance, the logit model has the highest cross-validation error rate (an average of 14.6%), while RF is the best of the algorithms, with a mean error rate of 5.2%. RF outperforms logit, LASSO, and SVM in accuracy, recall, specificity, precision, negative predictive value, F1 score, and metrics throughout the years analyzed, suggesting that this method can potentially contribute to the development of a high-frequency poverty measure for Mexico. The RF results show that the most important variables for multidimensional poverty prediction are the indicator variable of per capita real labor income below the minimum welfare line, household real labor income, household size, state of residence, and the binary variables for rural vs. urban localities and social security affiliation.

Previous studies employing similar methods analyze poverty solely in its monetary dimension and are focused on out-of-sample performance evaluation. Using data from Albania, Ethiopia, Malawi, Rwanda, Tanzania, and Uganda, Sohnesen and Stender (2016) compare RF out-of-sample performance against the usual poverty prediction methods, and find the lowest average MSE at the national level (1.71%) and at urban-rural levels (2.58%). However, they note that none of the methods is consistently accurate enough to predict poverty over time. Thoplan (2014) uses the 2000 census of Mauritius to train RF for income poverty classification and finds an error rate between 15% and 20%. In contrast to the results for Mexico, he finds that the number of hours worked, age, education, and gender are the most important variables for predicting poverty in that country. McBride and Nichols (2018) present evidence that ML ensemble methods can improve out-of-sample poverty detection by 2.43% for East Timor and 2.06% for Malawi over traditional proxy means tests (PMT).[4]

Data regarding such methods in Mexico is limited. Babenko *et al.* (2017) develop an income poverty map for Mexico based on an ML algorithm trained with Planet and DigitalGlobe imagery and the 2014 MCS-ENIGH. According to their estimates, most of the poverty rates of central and southern municipalities range between 48% and 70%, while in the northern region, there are more municipalities with rates between 20% and 30%. Satellite images, however, do not present

[4] One example of international interest in innovative approaches to machine learning and poverty targeting was a global data science competition, hosted by Driven Data and the World Bank at the beginning of 2018, called "Pover-T Tests: Predicting Poverty", to develop ML models for the prediction of household poverty. See https://www.drivendata.org/competitions/50/worldbank-poverty-prediction/page/99/.

enough variation from one quarter to another to develop a high-frequency measure. More recently, the EQUIDE Institute published a report outlining the economic impact of the COVID-19 pandemic in Mexico, based on its ENCOVID-19 survey. Their simulations suggest that poverty increased to 52% in May 2020, 10 pp above the official 2018 CONEVAL level, and that it rose to 54% in March 2021 (EQUIDE, 2021). The present study differs from these approaches in three key aspects: the source of information, the core methodology, and the scope of the analysis. It contributes to the growing literature on ML applications in economics by implementing novel algorithms in multidimensional poverty classification tasks in order to provide timely estimates for anti-poverty policymaking.

The structure of the paper is as follows. Section 2 presents a brief description of the ML algorithms. Section 3 describes the training and prediction data and the selected variables. Section 4 presents the optimal models, an overview of the impact of the pandemic, and a comparative performance assessment of the algorithms. Concluding remarks are offered in section 5. The appendix provides descriptive statistics, metrics, and ML poverty estimates.

2. Machine learning classification algorithms

The challenge of developing a high-frequency poverty measurement for Mexico can be addressed using the supervised ML subfield, and in particular its classification algorithms. This approach seeks the relationship between the objective variable Y and its observable characteristics $X_1, ..., X_P$ in order to find generalizable patterns and accurately predict future observations. Unlike standard methods that aim to make a good estimate of some parameter β with certain properties based on a functional form between Y and $X_1, ..., X_P$ ML classification algorithms focus on the exact estimation of the objective variable ($\hat{Y}$). In a word, supervised ML belongs in the part of the toolbox marked as $\hat{Y}$ rather than in the more familiar $\hat{\beta}$ compartment (Mullainathan and Spiess, 2017).

This section presents three popular algorithms in the ML literature: the LASSO logistic regression, the random forest method, and support vector machines, and it briefly describes their details, interpretability, and tuning parameters. This description follows James *et al.* (2013), Hastie *et al.* (2009), and Lantz (2015).

2.1 *The LASSO logistic regression model*

The least absolute shrinkage and selection operator, better known by its acronym LASSO, was designed for linear regression models, but its concept of shrinkage has been generalized to classification tasks through a combined framework with logistic regression that also performs variable selection. Formally speaking, consider a labeled dataset $\{(x_1, y_1), .., (x_n, y_n)\}$, where $x_i^T = (x_{i1}, ..., x_{ip})$ is the vector of the ith individual's features (also called predictors), and y_i is its corresponding label. We will assume that the objective (or response) variable Y is categorical and can take only two possible values (or classes), for instance, $y_i = 1$ if the ith individual is poor and $y_i = 0$ otherwise.[5]

The coefficients for the LASSO logistic regression are then obtained by solving the following penalized log-likelihood problem:

$$\underset{\beta,\beta_0}{\text{Min}} \; -\left\{\frac{1}{n}\sum_{i=1}^{n} y_i(x_i^T\beta + \beta_0) - \log(1 + \exp(x_i^T\beta + \beta_0))\right\} + \lambda\sum_{i=1}^{p}|\beta_i| \tag{1}$$

where the non-negative tuning parameter λ is chosen via cross-validation. The fitted probabilities $\hat{P}(Y = 1|X = x_i)$ are then computed with the estimated coefficients and the logistic transformation as in a logit model, and a simple classifier can be set according to the maximum probability criterion, that is, by assigning an observation x_i to the kth class if $\hat{P}(Y = k|X = x_i) \geq \hat{P}(Y = k'|X = x_i)$ for $k \neq k' \in \{0, 1\}$.

Note that in expression (1), λ controls the penalty in the optimization problem, and when it is equal to zero, the result gives the usual logit fit. This model has also the advantage of shrinking some coefficient estimates to zero, yielding a parsimonious model that is easier to interpret. Unlike traditional LASSO, its predicted probabilities do not fall outside the unit interval, which makes it a useful method for classification.

[5] See Friedman *et al.* (2010) for a more general version of the model considering $K > 2$ different classes.

 https://doi.org/10.24201/ee.v38i1.435

2.2 *Random forest*

The RF method is a well-known tree-based approach that aggregates a collection of classification trees to provide an accurate prediction for a new observation.[6] Each tree $T(x;\theta)$ in the forest segments the predictor space $X_1, ..., X_P$ into J mutually disjoint regions (or nodes) $R_1, ..., R_J$ and classifies new observations according to the most frequent class k_j into the region in which they belong.[7] Importantly, every region R_j depends on the minimum number of observations in each terminal node $n_{\min}$ or leaf), which is usually determined via cross-validation. Figure 2 depicts a classification tree with five terminal nodes in a three-dimensional feature space, and a binary response. Note that two advantages of classification trees are their graphic interpretability, which makes them easy to explain and understand compared to other black-box algorithms, and their good performance with highly non-linear decision boundaries. However, the method is very sensitive, since a small change in the training data can produce significant changes in the final tree. In other words, classification trees suffer from high variance.

Fortunately, this problem can be addressed by including a large number of trees in a single model. The RF classifier takes this idea and develops a powerful prediction algorithm with lower variance, in which similar training data yields almost identical results. Intuitively speaking, the RF iterative algorithm decorrelates its classification trees in two main steps. First, it draws different bootstrap samples from the training data to grow each tree, and second, it considers only a subset of variables to optimally split the final nodes.

We thus obtain a collection of decorrelated classification trees $\{T(x;\theta_b)\}_{b=1}^{B}$, best known as a RF, which takes the majority vote of its trees as the class prediction for a new observation x.[8] There is a trade-off between RF's variance and its interpretability: a decrease in variance is gained at the expense of an intuitive graphical representation of classification trees. The method has three main tuning

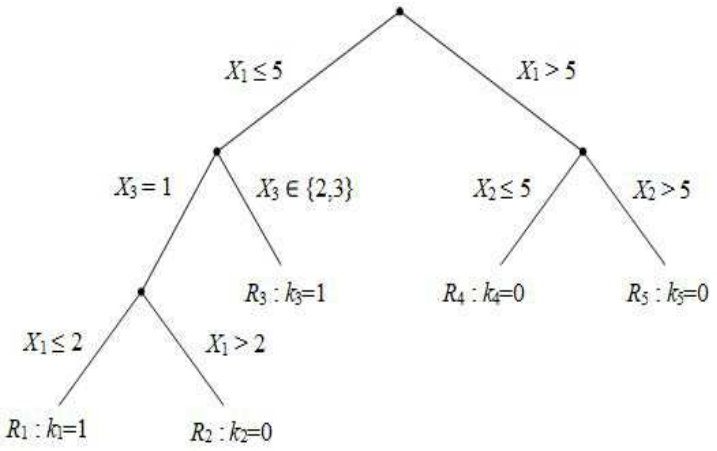
[6] For further details, see Breiman (2001), the seminal work on the random forest method.

[7] A classification tree can be expressed as $T(x;\theta) = \sum_{j=1}^{J} k_j I(x \in R_j)$, where $\theta = \{R_j, k_j\}_{j=1}^{J}$ is the parameter that gathers all the relevant information of the process of building the tree.

[8] A different threshold can be set instead of the majority vote, depending on the type of misclassification to be minimized.

parameters: the number of trees in the forest B, the number of variables considered at each split m, and the minimum node size on every leaf $n_{\min}$. In practice, we use a value of B sufficiently large such that the error settles down ($B \geq 500$ is generally sufficient), and the trees are grown deep (generally with $n_{\min} = 1$; James *et al.*, 2013). The tuning process is therefore centered around the parameter m, which controls the similarity among trees and prevents them from being highly correlated, resulting in a more reliable classifier.

Figure 2
Example of a classification tree $T(x;\theta)$ diagram



Note: In this case the tree has three predictors, X_1, X_2, X_3, two possible classes, $k \in \{1, 0\}$, and a partition of five regions $\{R_1, ..., R_5\}$.

The RF building process allows for simultaneous estimation of the error rate without the need to perform an extra validation process. Approximately one-third of the observations are left out in every bootstrap iteration; for every observation x_i it is thus possible to predict a response $\hat{y}_i$ using the majority vote of trees which did not include that observation in their training data. Then, the estimated misclassification rate, known as the out-of-bag error rate, would be given by $ER = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$. In addition, the RF method quantifies importance of variable to the predictive power of the model using two different measures: the mean decrease in accuracy (MDA), and the mean decrease in Gini index (MDGI). The former results from averaging, for all B trees, the total decrease in accuracy due to a random

permutation of the values of a given feature on the omitted observations in the corresponding trees, while MDGI is the average, for all B trees, of the total reduction in the Gini index over all splits of a given predictor.

2.3 *Support vector machines*

Another supervised learning method is support vector machines (SVM), whose main idea relies on establishing non-linear decision boundaries in the original feature space $X_1, ..., X_P$ through linear boundaries in a transformed and augmented space.[9] Like RF, SVM is considered a "black-box" algorithm because of its complex internal process. The SVM classifier exploits the idea of the natural separation produced by a hyperplane $H(\beta, \beta_0) = \{x \in \mathbb{R}^p : x^T\beta + \beta_0 = 0\}$, where new observations are classified according to the side on which they fall.

More specifically, a kernel function $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$ is chosen to characterize the transformation $h = (h_1, ..., h_m) : \mathbb{R}^p \to \mathbb{R}^m$ that will expand the original p-feature space into a higher-dimensional one $(m \geq p)$, where K and h are related by $K(x, x') = h(x)^T h(x), \ \forall x, x' \in \mathbb{R}^p$.[10] After obtaining the basis functions $h_j, \ j = 1, ..., m$, SVM finds the optimal hyperplane using the transformed observations as training data: $\{(h(x_1), y_1), .., (h(x_n), y_n)\}$. An important characteristic of this method is that it depends only on those observations that are closest to the hyperplane and lie exactly on its margin; these are known as support vectors. Figure 3 shows an example of linearly separable training data in a transformed bidimensional feature space, and its corresponding optimal hyperplane and support vectors.

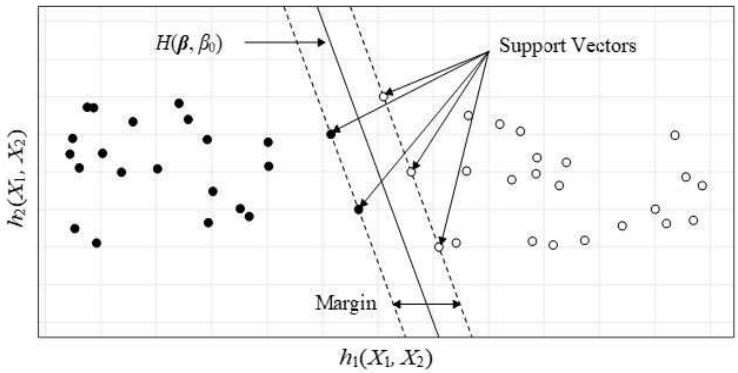
Note that, depending on the kernel used, there might be different support vector machines even with the same data. Table 1 shows some kernels popular in the ML literature. As expected, the radial basis and dth-degree polynomial kernels result in more flexible decision boundaries than the linear kernel. In addition, SVM has a cost parameter C, which controls the smoothness of the hyperplane boundary, and together with the kernel's parameters is selected through cross-validation using a tuning grid.

---

[9] Additional details about SVM can be found in the influential work of Cortes and Vapnik (1995).

[10] Kernel functions, rather than other transformations, are used to define the basis function h owing to computational issues.

Figure 3
Optimal hyperplane and support vectors



Note: Linearly separable training data, in a transformed bidimensional space.
Source: Lantz (2015).

Table 1
Kernel functions and tuning parameters

Kernel	$K(x, x')$	*Parameters*
Linear	$x^T x'$	None
dth-degree polynomial	$\left(\gamma x^T x' + r\right)^d$	$d \in \mathbb{Z}_+, r, \gamma > 0$
Radial basis	$\exp\left(-\gamma \mid\mid x - x' \mid\mid^2\right)$	$\gamma > 0$

Source: Hastie *et al.* (2009).

Nonetheless, SVM suffers from a major drawback: it can be very slow to train, particularly if the input dataset has a large number of features or observations (Lantz, 2015). Steinwart and Thomann (2017) suggest an alternative approach, in which the feature space is split into spatial cells where local SVMs with the same kernel are trained, thus considerably decreasing runtime with large samples. Furthermore, cross-validation is performed on every cell, resulting in the same number of optimal tuning parameters as cells. Then, to classify a new observation x, the SVM of the cell to which x belongs is used. According to the authors, the solution of the cell approach is similar to that of SVM.

3. Data

A natural choice for the labeled dataset, or training data, described in the previous section is the MCS-ENIGH (Socioeconomic Conditions Module of the National Survey of Household Income and Expenditures), in which CONEVAL identifies and labels multidimensional poverty at the individual level, according to its official methodology. Biennial survey data from 2008 to 2018 is used to build six training datasets, one for each MCS-ENIGH, resulting in six biennial windows to classify new observations.[11] It should be noted that since 2015, there have been methodological changes in the MCS that do not allow historical comparison with its previous versions.[12] For 2016 and 2018, data is therefore taken from the Statistical Model for the Continuity of Socioeconomic Conditions Module (MEC) of the MCS-ENIGH, provided by the National Institute of Statistics and Geography (*Instituto Nacional de Geografía y Estadística*, INEGI).

In order to obtain a higher-frequency poverty measure, the fitted models are applied to ENOE observations, the prediction data, corresponding to all quarters from 2008 to 2019. The publication history of the CONEVAL multidimensional poverty figures and the last quarter of the ENOE available at that time are used to delimit the prediction periods of each model depending on their training year, as shown in table 2. Thus, for instance, models trained on the 2008 MCS-ENIGH are used to predict the poverty status of all ENOE observations from the first quarter of 2008 until the last quarter available before the publication date of the 2010 multidimensional poverty measure, i.e., from ENOE 2008 QI to ENOE 2011 QI. In a similar way, models fitted with the 2010 training data predict the poverty status of all ENOE observations from the first quarter of 2010 to the last available quarter before the 2012 multidimensional poverty measurement is released, i.e., from 2010 QI to 2013 QI, and so on. Note that in this approach,

[11] The strategy of using biennial classification windows hinges on the assumption that the economic environment does not vary substantially in the short to medium term, so it is valid to predict the poverty status of an individual in all quarters of years t and $t+1$ with an algorithm trained in year t. A better approach would be to model and predict only within the same year, as suggested by Sohnesen and Stender (2016). However, given the publication frequency of the MCS-ENIGH, this approach is not feasible.

[12] One of the main changes was in the interviewers' method of collecting data on family income. See INEGI Press Release No. 286/16 for more details of the methodological changes in the 2015 MCS.

prediction periods are wider than biennial spans, which in turn yields intersections between consecutive periods. For an algorithm trained in year t, it is then convenient to define its ex-post estimates as its poverty predictions over quarters inside the biennial window $[t,\ t+1]$, while ex-ante estimates are defined as those outside this window. The latter are updated with their corresponding ex-post estimates, made by the same algorithm but trained with data from year $t+2$, when CONEVAL releases its next multidimensional poverty measure.

Table 2
Publication dates of poverty measures and latest ENOE

	2008	*2010*	*2012*	*2014*	*2016*	*2018*
Multidimensional poverty	Dec. 2009	Jul. 2011	Jul. 2013	Jul. 2015	Aug. 2017	Aug. 2019
Latest ENOE	2009 QIII	2011 QI	2013 QI	2015 QI	2017 QII	2019 QI

Note: "Latest ENOE" shows the last quarter of the survey that was available at the time each multidimensional poverty measure was published.

Source: CONEVAL and INEGI.

Because these surveys are designed for different purposes, they have many variables that are not compatible. The first step is thus to find comparable questions and variables that will have the same features in both the training and prediction datasets.

3.1 *Selected variables*

Comparable features of the MCS-ENIGH and the ENOE are divided into four sets of variables. The first is the sociodemographic set, which consists of state of residence, binary variables for rural/urban and gender, and continuous variables for age and household size. The second is the set of economic characteristics, which includes a categorical variable for individual economic status, continuous variables for household hours of work per week and real labor income per month, and an indicator variable that equals one if household per capita real labor income is below the minimum welfare line (mwl), and zero oth-

erwise.[13] The third set consists of educational variables, formed by an indicator variable for not in school, a categorical variable for educational level, and an indicator variable for educational lag. Finally, there is a set of binary variables for access to public health care (*Instituto Mexicano del Seguro Social*, IMSS; and *Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado*, ISSSTE).

Table 3
Selected and input variables between training and prediction data

Description	*Name*	*MCS-ENIGH*	*ENOE*
Sociodemographic features			
State	state	ent	ent
Dummy for rural localities	rururb	rururb	rururb
Gender	gender	sexo	sex
Age	age	edad	eda
Household size	hsize	built	built
Economic status			
Economically active population	eap	pea	clase1/clase2
Household hours worked per week	hhwork	htrab	hrsocup
Household labor income per month	lab_inc	ing_lab	p6b2/p6c
Households with per capita labor income less than the mwl	pob	ing_lab	p6b2/p6c
Education			
Not in school	no_school	inas_esc	cs_p17
Educational level	ed_lev	niv_ed	cs_p13_1
Educational lag	ed_back	ic_rezedu	cs_p17/cs_p13_1 cs_p13_2/cs_p15

[13] MCS-ENIGH income uses the CONEVAL deflation process in the computation of multidimensional poverty, whereas ENOE income uses the National Consumer Price Index (INPC) for August of each year of the ENIGH. That is, the ENOE income in the first prediction period [2008 QI, 2011 QI], which is linked to the 2008 training data, is thus deflated by the INPC for August 2008. Income in the second span [2010 QI, 2013 QI], which is linked to the 2010 training data, is deflated by the INPC for August 2010, and so on.

Table 3
(Continued)

Description	*Name*	*MCS-ENIGH*	*ENOE*
Health care			
IMSS affiliation	imss	serv_sal	imssisssste
ISSSTE affiliation	issste	serv_sal	imssisssste
No health insurance	no_hserv	serv_sal	imssisssste

Note: The third and fourth columns show the names of the input variables as they appear in the original surveys.

Source: Author's elaboration using documentation from MCS-ENIGH and ENOE.

Table 3 summarizes these predictors and shows the input variables from MCS-ENIGH and ENOE that are used to create them. The ENOE input variables are taken directly from INEGI releases, while the MCS-ENIGH input variables, with the exception of household size and hours worked per week, are constructed following the official CONEVAL methodology for the multidimensional poverty measurement.[14] Although the number of predictors is relatively small, it is worth noting that they not only describe the dimension of household income, but also individual educational trajectory (especially educational lag) and household access to health care. The collected predictors consider three out of nine of the dimensions established in the General Law of Social Development for multidimensional poverty to expand the one-dimensional approach used to identify labor poverty. The objective variable for poverty is constructed in the same way as in CONEVAL (2014) and included in the training data. This variable identifies poor and non-poor individuals according to the multidimensional poverty measurement methodology, and is the dependent variable in the prediction data since it is not possible to obtain it deterministically from the ENOE.

3.2 *Training and prediction data*

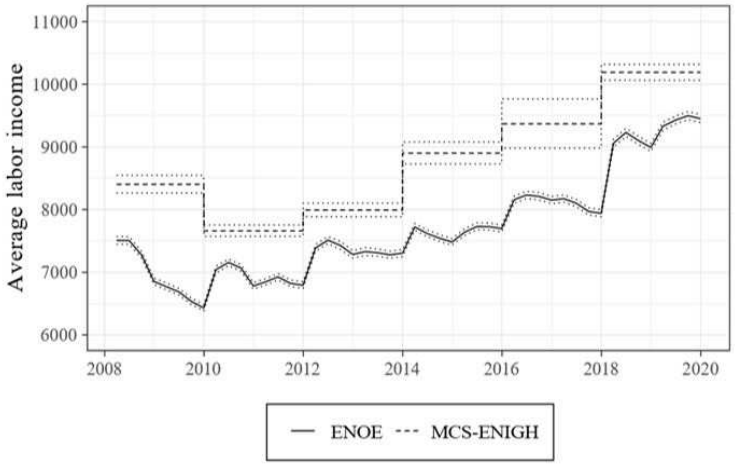
Some ML algorithms display high variance: their output may vary widely with small changes in the training data. They may produce

[14] CONEVAL's computation programs are available at: https://www.coneval.org.mx/Medicion/MP/Paginas/Programas_BD_08_10_12_14_16_18.aspx.

 https://doi.org/10.24201/ee.v38i1.435

unexpected estimates if the distribution of characteristics differs substantially from one survey to the other. It is thus important to identify similarities and differences in the proportions and dynamics of the series in both training and prediction data.

Figure 4
Average labor income of Mexican households



Note: 95% confidence intervals are depicted with dotted lines.

Source: Author's calculations using deflated income data from MCS-ENIGH and ENOE, from 2008 to 2019, without expansion factors.

Descriptive statistics for both data sources are shown in Appendix A.[15] Tables A.1 and A.3 to A.8 show descriptive statistics for the household-level variables from training and prediction data. Mean household size and average hours worked per week have remained relatively stable over time, but only household size is comparable between surveys; households in the prediction data work an average of eleven hours less per week than those in the training data. Mean labor income is also not comparable between datasets (and thus the proportion of households with per capita labor income below the minimum welfare line is also not comparable). Figure 4 shows the contrast between average real labor income of Mexican households in training and prediction data over time; the income in the training

[15] Means and standard deviations are computed without using population expansion factors, in order to compare the raw data.

data is always higher. The confidence intervals suggest the existence of statistically significant differences between the biennial means, although they exhibit similarly increasing behavior since 2010. Table A.15 summarizes the *t*-test results for this variable, where the null hypothesis of mean equal income between surveys within their corresponding biennial window (from 2008 to 2009, from 2010 to 2011, etc.) is rejected at all significance levels, meaning that the predictions with ML algorithms will be affected to some extent by the income patterns of the training data. Additionally, the methodological changes inherited by the 2016 MEC bring about a noteworthy increase in labor income dispersion over previous years, mainly due to the attempt by INEGI to improve its household income data by means of stricter capture and verification field criteria; these changes affect the income distribution in the 2016 training data. This pattern is not displayed in the 2018 MEC.

Similarly, tables A.2 and A.9 to A.14 show individual-level descriptive statistics from the training and prediction data. It can be seen that most variables display steady behavior that does not differ substantially over time or across datasets. However, the fraction of individuals in rural localities in the training data is, on average, 12 pp greater, which in turn means a smaller number of people affiliated with IMSS and a higher level of educational lag. There is a significant decrease in the training data on the number of individuals who do not have any access to health care services, from 34.8% in 2008 to 14% in 2018; in the prediction data this variable hovers around 25%. Finally, the objective variable of multidimensional poverty is balanced over time: neither poor nor non-poor individuals exceed 60% of the total number of observations in the training datasets, and the number remains very close to 44%. These numbers are equivalent to the official poverty estimates using the corresponding expansion factors.

4. Results

Besides the labor poverty rate benchmark, the traditional econometric toolbox provides several other methods to tackle the poverty classification problem. In order to compare the ML algorithms with a traditional approach, the following logistic regression specification is used as the baseline model for each training dataset:

$$P\left(Y_i = 1 \,|X_i\,\right) = \frac{e^{\alpha+\beta^T X_i}}{1 + e^{\alpha+\beta^T X_i}} \tag{2}$$

where the dependent variable Y_i is the multidimensional poverty indicator variable that takes the value of 1 if the ith individual is poor and 0 otherwise, and X_i is the ith vector of predictors with all the selected features of section 3.1 ($state_i, rururb_i, sex_i$, and so on).

Table 4 reports the logit output for every training dataset. Most of the coefficients keep their sign over time and are statistically significant at the 1% level. An advantage of this model is that it allows for inferences about variations in the probability of being poor if a predictor changes. For instance, it can be inferred from the sign of the coefficients of the continuous variables that a *ceteris paribus* increase in age or household labor income would decrease the probability of being poor, whereas rising household size or hours of work would increase that probability. In a slightly more elaborate way, one can show that living in a rural locality, being female, having completed at least junior high school, or having an IMSS or ISSSTE affiliation lead to a lower probability of being poor with respect to the reference level, whereas living in Chiapas, being part of the unemployed economically active population, not attending school, having an educational lag, or not having health insurance would increase it. Table B.1 shows the logit average marginal effects (AME) on the probability of being poor; interestingly, AMEs for the three poorest states in 2018 (Oaxaca, Guerrero, and Chiapas) have increased an average of almost four times from 2008 to 2018, reflecting the low efficiency of policies targeting poverty in those states.[16]

The logit is a probabilistic model that is not designed for classification tasks per se. Hence, it is necessary to build a classifier based on its predicted probabilities. The simplest way to do this is to set a 50% cutoff (c) and classify an observation as poor if its predicted probability of belonging to this class exceeds this threshold. Also, c can be seen as a tuning parameter, and its value can thus be determined with cross-validation. In this application, the optimal cutoff c_t for the logit model fitted with training data for year t is determined through the minimization of the mean square error between the ex-post poverty estimates that are available at the time of the publication of the multidimensional poverty rate for year t and the corresponding official multidimensional poverty level. Table 4 reports these optimal cutoffs, where there is a 58% threshold over almost all

[16] The AME of non-independent predictors, such as the indicator variable of per capita labor income below the mwl, should be taken with a grain of salt, given that a shift on this level is the result of a change in labor income or household size, invalidating the *ceteris paribus* assumption.

Table 4
Logistic regression results

	Dependent variable: Probability of being poor					
	2008	*2010*	*2012*	*2014*	*2016*	*2018*
Nuevo León	-0.301***	-0.181***	0.096	0.036	0.180***	0.219***
	(0.061)	(0.058)	(0.059)	(0.058)	(0.049)	(0.051)
Baja California Sur	-0.090	0.008	0.092	0.445***	0.258***	0.389***
	(0.061)	(0.057)	(0.058)	(0.056)	(0.052)	(0.054)
Coahuila	-0.118**	-0.123**	-0.139**	0.112**	0.234***	0.205***
	(0.056)	(0.054)	(0.054)	(0.055)	(0.045)	(0.047)
Oaxaca	0.201***	0.762***	0.577***	0.694***	1.181***	1.176***
	(0.055)	(0.054)	(0.053)	(0.053)	(0.051)	(0.050)
Guerrero	0.396***	0.365***	0.644***	0.457***	0.876***	1.002***
	(0.058)	(0.055)	(0.056)	(0.053)	(0.052)	(0.053)
Chiapas	0.530***	0.779***	0.787***	0.881***	1.332***	1.117***
	(0.058)	(0.051)	(0.055)	(0.055)	(0.054)	(0.055)
Rural localities	-1.807***	-1.527***	-1.430***	-1.460***	-1.531***	-1.561***
	(0.018)	(0.017)	(0.018)	(0.018)	(0.015)	(0.015)
Women	-0.067***	-0.017	-0.019	-0.051***	-0.029**	-0.054***
	(0.014)	(0.014)	(0.014)	(0.014)	(0.013)	(0.013)
Age	-0.014***	-0.013***	-0.013***	-0.013***	-0.013***	-0.013***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.0005)
Household size	0.906***	0.858***	0.794***	0.784***	0.830***	0.892***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Employed EAP	0.216***	0.288***	0.284***	0.249***	0.258***	0.293***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.017)	(0.017)
Unemployed EAP	0.249***	0.399***	0.389***	0.460***	0.395***	0.394***
	(0.047)	(0.042)	(0.045)	(0.047)	(0.053)	(0.052)

Table 4
(Continued)

	Dependent variable: Probability of being poor					
	2008	*2010*	*2012*	*2014*	*2016*	*2018*
EAP under 15	-0.300***	-0.316***	-0.309***	-0.273***	-0.238***	-0.293***
	(0.034)	(0.034)	(0.036)	(0.035)	(0.033)	(0.032)
Hours worked per week	0.005***	0.004***	0.002***	0.004***	0.005***	0.005***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Labor income per month	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.0005***
	($5x10^{-6}$)	($4x10^{-6}$)	($4x10^{-6}$)	($4x10^{-6}$)	($4x10^{-6}$)	($3x10^{-6}$)
Households with per capita labor	-0.513***	-0.410***	-0.218***	-0.202***	-0.224***	-0.246***
income lower than the mwl	(0.021)	(0.021)	(0.021)	(0.021)	(0.019)	(0.019)
Not in school	0.208***	0.126***	0.136***	0.120***	0.110***	0.100***
	(0.021)	(0.020)	(0.021)	(0.021)	(0.019)	(0.019)
Completed elementary school	-0.456***	-0.233***	-0.263***	-0.195***	-0.180***	-0.185***
	(0.022)	(0.022)	(0.023)	(0.022)	(0.020)	(0.020)
Completed junior high school	-0.765***	-0.454***	-0.458***	-0.309***	-0.278***	-0.265***
	(0.028)	(0.028)	(0.029)	(0.029)	(0.026)	(0.026)
Educational lag	0.455***	0.720***	0.693***	0.833***	0.795***	0.787***
	(0.025)	(0.025)	(0.026)	(0.026)	(0.023)	(0.023)
IMSS affiliation	-1.419***	-1.593***	-1.569***	-1.596***	-1.482***	-1.522***
	(0.019)	(0.018)	(0.018)	(0.018)	(0.016)	(0.016)
ISSSTE affiliation	-2.088***	-2.309***	-2.463***	-2.431***	-2.503***	-2.542***
	(0.040)	(0.039)	(0.044)	(0.045)	(0.045)	(0.043)
No health insurance	0.163***	0.124***	0.124***	0.173***	0.165***	0.035**
	(0.017)	(0.016)	(0.018)	(0.018)	(0.018)	(0.018)

Table 4
(Continued)

	Dependent variable: Probability of being poor					
	2008	*2010*	*2012*	*2014*	*2016*	*2018*
Constant	0.894***	0.594***	0.648***	0.338***	0.046	0.027
	(0.062)	(0.061)	(0.062)	(0.061)	(0.056)	(0.057)
Optimal cutoff	0.58	0.58	0.58	0.58	0.56	0.58
Error rate	0.135	0.138	0.148	0.151	0.153	0.153

Note: *p<0.1; **p<0.05; ***p<0.01. Standard errors are shown in parentheses. 10-fold CV error rates are computed using their corresponding optimal cutoff. The first three states are below the 10th percentile on the 2018 national poverty scale, while the last three are above the 90th percentile. The remaining states coefficients are not displayed. The reference levels of categorical variables are Aguascalientes, economically inactive population, and incomplete elementary school or less.

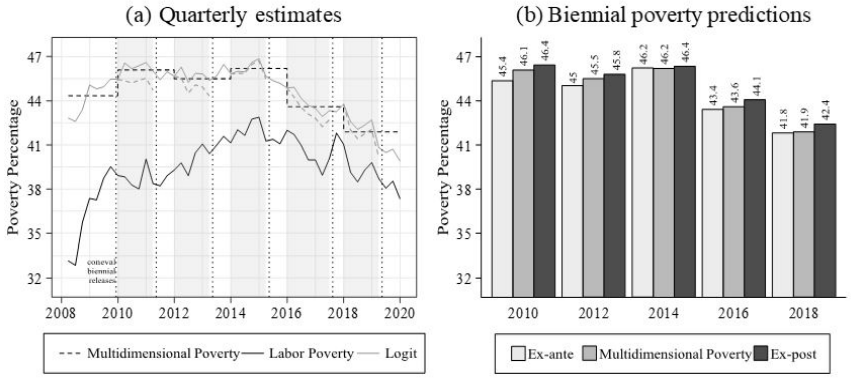
Source: Author's calculations using the MCS-ENIGH from 2008 to 2018, and the stats package (version 3.5.0) in R.

 https://doi.org/10.24201/ee.v38i1.435

of the six training data years.[17]

Figure 5 shows the resulting series of the benchmark model. Note that in panel (a), both ex-ante and ex-post logit poverty rates correctly reflect the downturns and upturns of the multidimensional rate on a quarterly basis, and in general, they are also closer to it than labor poverty. The ex-post logit estimates also follow the increasing trend of labor poverty from 2008 to 2010, where the labor poverty increase of 6.4% during the world financial crisis of 2008-2009 is more than twice that of the logit rate of 2.7%. Unsurprisingly, from 2010 to 2015, the estimates of the baseline model are guided by the patterns of the training data, and the ex-post estimates hover around 46%, while labor poverty increases to 43%. Both series show a similar decline from 2015. More importantly, the ex-ante logit estimates, which can be considered as quarterly predictions for the next multidimensional poverty level, show small differences over time (an average of 0.67 pp) with their respective ex-post updates.

Figure 5
Multidimensional, labor, and logit poverty rates



Source: Author's calculations using the MCS-ENIGH and ENOE from the first quarter of 2008 to the last quarter of 2019, and the stats package (version 3.5.0) in R.

Figure 5, panel (b), shows the comparison between the biennial multidimensional poverty levels and the quarterly averages of the ex-ante and ex-post logit estimates within their corresponding years.

[17] By increasing the cutoff from 50% to 58%, the poverty condition becomes more stringent and the number of people misclassified as poor (Type I error) is reduced, while the opposite occurs with Type 2 error.

The ex-ante predictions slightly underestimate the official poverty rate, while the ex-post estimates slightly overestimate it, underlining the consistent behavior of the baseline estimates. The advantage of this approach of having ex-ante estimates is that a multidimensional poverty estimate can be forecast almost six months before the official biennial publication. For instance, it would have been known since February 2019 that the 2018 multidimensional poverty rate, which was announced in August 2019, would be close to 41.8%. Finally, the error rate of the logit increases from 13.5% in 2008 to 15.3% in 2018, with an average rate of 14.6%, as shown in table 4.

4.1 *Poverty estimates of the LASSO logistic model*

As described in section 2.1, the LASSO logistic model can handle several regressors in a logistic regression framework and select only a subset of them by shrinking the coefficients of some variables to zero. In this analysis, quadratic terms for continuous variables and interactions among categorical variables (except for the state of residence variable) are added to the baseline specification in (2), and the algorithm decides which variables to keep. The penalty term λ is determined using a tuning grid and 10-fold cross-validation.[18]

Table 5 presents a summary of the estimation results, where the algorithm drops an average of one-tenth of the variables each year. This amount is partially determined by the small value of the penalty parameter λ: the larger it is, the small the number of non-zero coefficients in the model. Although some coefficients are close to those in table 4, there is an average 1.8 pp gain in accuracy over the logit out-of-sample performance. Even if the interaction and quadratic terms are not included, the LASSO logistic model outperforms the logit by 0.7 pp, as shown in table B.2. In other words, adding more variables to the baseline specification in this algorithm to cover possible non-linearities yields a small improvement in the misclassification rate.

However, the LASSO logistic ex-post and ex-ante estimates provide closer approximations to the multidimensional poverty rate than labor poverty in the sample analyzed, as shown in figure 6, panel (a). Both poverty measures show an increasing trend from 2008 to 2010, where ex-post estimates increase by 2.3 pp during the world financial crisis. LASSO's multidimensional poverty estimates then hovers around 48%, while labor poverty continues to increase until 2015.

[18] The tuning grid used for the penalty term is:
$\lambda \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000\}$.

Table 5
LASSO logistic regression results

	Dependent variable: Probability of being poor					
	2008	*2010*	*2012*	*2014*	*2016*	*2018*
Nuevo León	-0.373	-0.268	0.000	-0.087	0.070	0.028
Baja California Sur	-0.056		0.081	0.320	0.045	0.200
Coahuila	-0.113	-0.175	-0.194	-0.048	0.085	0.005
Oaxaca	0.141	0.683	0.481	0.513	0.950	0.933
Guerrero	0.263	0.189	0.512	0.263	0.657	0.767
Chiapas	0.401	0.681	0.688	0.707	1.112	0.879
Rural localities	-2.060	-2.056	-1.835	-1.918	-1.719	-1.802
Women	0.095	0.166	0.166	0.105	0.116	0.037
Age	-0.006		0.001	0.004	0.004	0.003
Household size	0.986	0.942	0.893	0.912	0.933	0.973
Employed EAP		0.051				
Unemployed EAP	0.144	0.547	0.504	0.080	0.559	0.434
EAP under 15			-0.021			
Hours worked per week	0.012	0.010	0.007	0.008	0.011	0.011
Labor income per month	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
Households with per capita labor income lower than the mwl	-0.908	-0.923	-0.883	-0.853	-1.042	-1.129
Not in school	0.184	0.005	0.019	0.071	0.178	0.113
Completed elementary school	-0.186			-0.042	-0.085	
Completed junior high school	-0.476	-0.292	-0.315	-0.147	-0.271	-0.212
Educational lag	0.528	0.735	0.727	0.856	0.717	0.893

Table 5
(Continued)

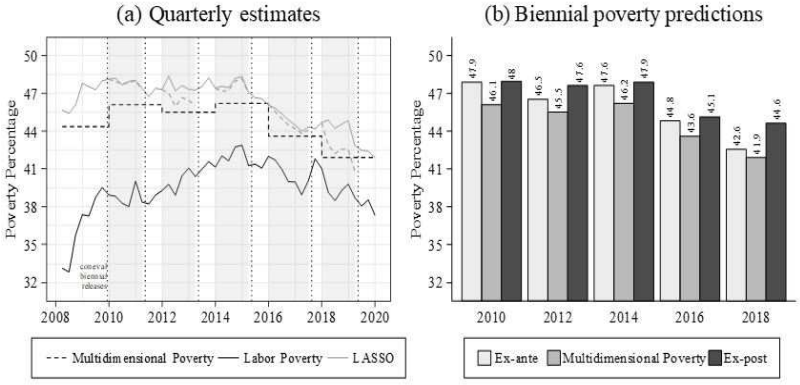
	Dependent variable: Probability of being poor					
	2008	*2010*	*2012*	*2014*	*2016*	*2018*
IMSS affiliation	-1.104	-1.712	-1.413	-1.257	-1.246	-1.232
ISSSTE affiliation	-1.461	-2.146	-2.224	-1.836	-2.331	-1.925
No health insurance	0.496	0.373	0.512	0.624	0.522	0.449
Constant	0.215	0.131	0.23	-0.126	-0.395	-0.222
Interaction terms	Yes	Yes	Yes	Yes	Yes	Yes
Non-zero coefficients (Total=127)	105	115	114	111	114	110
Penalty λ	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Error rate	0.117	0.121	0.129	0.133	0.137	0.134

Note: LASSO logistic models are estimated via penalized maximum likelihood. Empty spaces represent shrunk-to-zero coefficients. Independent variables include those in table 3, quadratic terms for age, household size, hours worked, and labor income, and interactions among categorical variables (except for state of residence variable). 10-fold CV error rates are computed according to the maximum probability criterion. The first three states are below the 10th percentile on the 2018 national poverty scale, while the last three are above the 90th percentile. The remaining coefficients are not displayed.

Source: Author's estimates using the MCS-ENIGH from 2008 to 2018, and the glmnet package (version 4.1-1) in R.

For the rest of the time window, the two rates show similar decreases, with the LASSO logistic estimates less volatile than those for labor poverty. The ex-ante and ex-post estimates are further apart during 2018, revealing some degree of sensitivity to the methodological changes in the 2016 and 2018 training data. Finally, according to the quarterly averages in figure 6, panel (b), ex-ante and ex-post measures overestimate the biennial multidimensional rate by an average of 1.2 pp and 2 pp, respectively.[19]

Figure 6
Multidimensional, labor, and LASSO logistic poverty rates



Note: Solid and dashed gray lines in panel (a) represent ex-post and ex-ante LASSO logistic estimates, respectively, while gray areas indicate where ex-ante estimates are seen before their updated ex-post version. Ex-ante and ex-post poverty rates in panel (b) are the averages of their respective quarterly estimates that fall within each publication year of multidimensional poverty. LASSO logistic estimates are computed using the maximum probability criterion.

Source: Author's calculations using the MCS-ENIGH and ENOE from the first quarter of 2008 to the last quarter of 2019, and the glmnet package (version 4.1-1) in R.

4.2 *Poverty estimates of random forest*

Similar to the baseline analysis, the random forest (RF) method uses all of the variables in table 3 to train the ML algorithm, and it uses the

[19] One of the reasons behind this gap is the 50% threshold used in the LASSO logistic classifier.

optimal cutoffs in table 4. As described in section 2.2, the RF method has three main tuning parameters: the number of trees in the forest (B), the number of possible predictors at each split (m), and the minimum size on every terminal node ($n_{\min}$). For this application, the parameters B and $n_{\min}$ are fixed at 500 and 1, respectively. RF calibration is thus focused on parameter m, for which a grid of possible values is built and tested using out-of-sample performance in order to choose the best model.[20] Optimal values of the tuning parameter m and out-of-bag error rates of their corresponding fitted models are shown in table 6. The optimal parameters are very close to the total number of predictors, and there is a significant improvement in RF of out-of-sample performance, exceeding the LASSO logistic and logit models by an average of 7.7 pp and 9.4 pp, respectively. This result suggests that decision boundaries are highly non-linear, and that the RF structure helps to assimilate them better than a traditional regression approach.

Table 6
Optimal parameters for the random forest analysis

	2008	*2010*	*2012*	*2014*	*2016*	*2018*
m	13	13	13	13	13	13
Error rate	0.048	0.049	0.052	0.055	0.052	0.056

Note: Number m of optimal predictors for each split in the RF classification trees. Out-of-bag error rates are computed using the optimal cutoffs for the benchmark model.

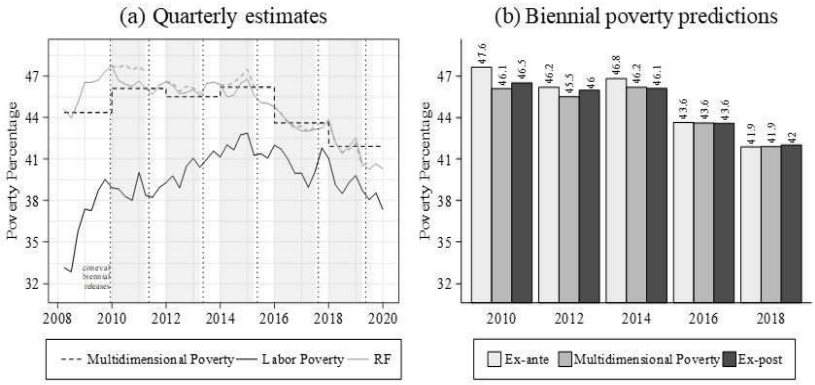
Source: Author's estimates using the MCS-ENIGH for 2008 to 2018, and the randomForest package (version 4.6-14) in R.

The RF quarterly poverty rate is depicted in figure 7, panel (a), which shows ex-post levels that are close to the official poverty rate. The increasing trend during the world financial crisis shows a jump of 2.8 pp, a magnitude similar to that of the benchmark model. From 2010 to 2015, it hovers around 46%, decreasing toward the end of the period. The RF ex-ante quarterly predictions are close to their corresponding ex-post updates, with an average absolute deviation of 0.5 pp. The quarterly averages in panel (b) show an interesting result:

[20] The best model is chosen according to the minimum out-of-bag error rate. The tuning grid used is $m \in \{2, 3, 4, 6, 9, 13, 15\}$.

the RF ex-post estimates generally improve their ex-ante poverty predictions, in contrast to the LASSO logistic estimates. Even in 2016 and 2018, RF adjusts better to the methodological changes in the training data than the logit model.

Figure 7
Multidimensional, labor, and RF poverty rates



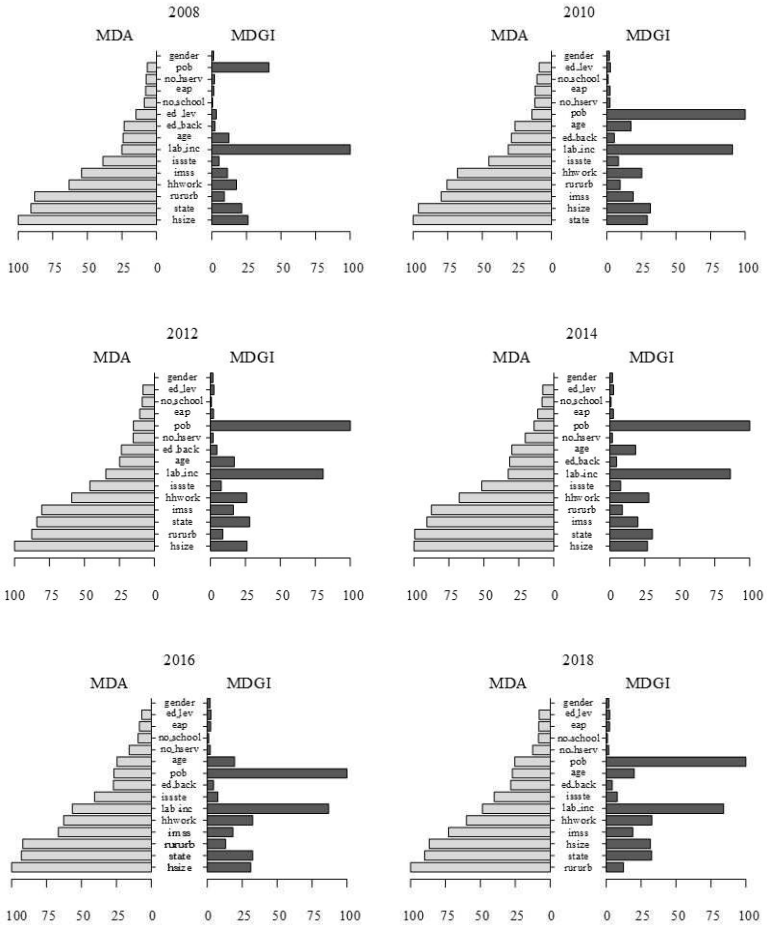
Note: Solid and dashed gray lines in panel (a) represent ex-post and ex-ante RF estimates, respectively, while gray areas are spans where ex-ante estimates are seen before their updated ex-post version. Ex-ante and ex-post poverty rates in panel (b) are the averages of their respective quarterly estimates that fall within each publication year of multidimensional poverty. RF estimates are computed using the optimal cutoffs for the benchmark model, in table 4.

Source: Author's calculations using the MCS-ENIGH and ENOE from the first quarter of 2008 to the last quarter of 2019, and the randomForest package (version 4.6-14) in R.

In spite of being a black-box algorithm, RF offers a picture of what is happening inside through the mean decrease in accuracy (MDA) and the mean decrease in Gini index (MDGI) as measures of variable importance. These measures are shown in relative terms in figure 8. According to the MDGI metric, per capita real labor income below the mwl (*pob*) and labor income (*lab_inc*) are the most important variables over the six years. For MDA, the four most important variables are household size (*hsize*), state of residence (*state*), the indicator variable of rural localities (*rururb*), and IMSS affiliation (*imss*). Gender, educational level (*ed_lev*), not in school (*no_school*), no health insurance (*no_hserv*), and economic status (*eap*) are the variables

with least importance over the six years. Surprisingly, access to social security has greater importance for the predictive power of the algorithm than educational level or school attendance, highlighting the critical role of formal employment in predicting poverty.

Figure 8
Variable importance for the random forest method



Note: Both mean decrease in accuracy (MDA) and mean decrease in Gini index (MDGI) are expressed relative to their corresponding largest measure and sorted by MDA, for each year of training data.

Source: Author's calculations using the MCS-ENIGH for 2008 to 2018, and the randomForest package (version 4.6-14) in R.

 https://doi.org/10.24201/ee.v38i1.435

4.3 *Poverty estimates of support vector machines*

Like the RF method, SVM analysis uses all the features of table 3. As described in section 2.3, the training times of local SVMs are considerably less expensive with large samples than training a single SVM, so this exercise, with approximately 230 000 training observations per year, uses the alternative SVM cell approach of Steinwart and Thomann (2017). The radial basis transformation is employed as the default kernel function, leading to two tuning parameters, C and λ, for which a tuning grid is built and tested, using 10-fold cross-validation to pick the best model for every cell.[21]

Table 7
Optimal parameters for the support vector machine cell approach

	2008	*2010*	*2012*	*2014*	*2016*	*2018*
No. of cells	186	170	174	173	211	215
Mean size	1,264	1,385	1,222	1,250	1,221	1,251
γ_{mode}	0.25	0.04	0.001	0.25	0.001	0.04
C_{mode}	100,000	100,000	100,000	100,000	100,000	100,000
Average support vectors	503	565	595	567	626	545
Mean error rate	0.119	0.115	0.121	0.134	0.15	0.123

Note: Number of cells used and their average size, along with their radial kernel optimal parameters (C and γ), modes, and average support vectors. The mean error rate corresponds to the average of the 10-fold CV error rates over all cells.

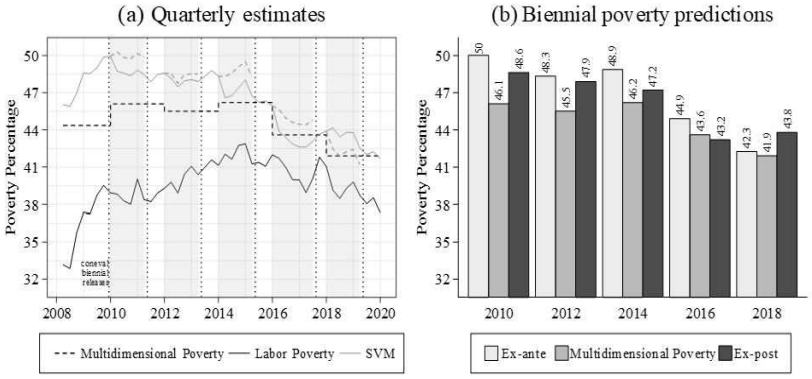
Source: Author's estimates using the MCS-ENIGH for 2008 to 2018, and the liquidSVM package (version 1.2.1) in R.

Table 7 shows summary results of the SVM cell calibration. There are more cells in 2016 and 2018 than in other years, mainly due to the greater number of observations and, to some extent, to the difference in the distribution of training data. The data in 2010 yields only an increase in average cell size, without affecting the number of cells. For every year most cells pick a small optimal value for λ, which is related to classifiers with high variance given that their kernels would be

[21] The tuning grid used is $(\gamma, C) \in \{0.001, 0.01, 0.04, 0.25, 1, 4, 16, 100\} \times \{0.001, 0.01, 1, 10, 100, 1000, 10000, 100000\}$.

large, and it would thus have a greater impact on the decision function $h(x)^T\hat{\beta} + \hat{\beta}_0$; higher values of λ are related to more non-linear fits and more bias classifiers. The mode of the cost parameter C remains constant over time, with a value of 100 000. According to Hastie *et al.* (2009), a large value of C encourages a large value of $\|\beta\|$, which brings about a narrower margin and fewer support vectors, so the model tends to overfit the data. Roughly two-fifths of the observations are support vectors on every cell, where the results for 2016 suggest that the greater the dispersion in the training data, the more support vectors are needed. Finally, the mean error rate of the SVM cells is taken as the estimate for the model's out-of-sample performance, which is similar to the LASSO logistic performance, slightly better than the benchmark model, but not as good as RF.

Figure 9
Multidimensional, labor, and SVM poverty rates



Note: Solid and dashed gray lines in panel (a) represent ex-post and ex-ante SVM estimates, respectively, while gray areas indicate the spans of ex-ante estimates before their updated ex-post version. Ex-ante and ex-post poverty rates in panel (b) are the average of their respective quarterly estimates that fall within each publication year of multidimensional poverty.

Source: Author's calculations using the MCS-ENIGH and ENOE from the first quarter of 2008 to the last quarter of 2019, and the liquidSVM package (version 1.2.1) in R.

Figure 9 shows the poverty rates using the optimal models of SVM cells. At first glance, the ex-post curve of SVM in panel (a) overestimates the multidimensional poverty rate during the first seven years of analysis, showing a decreasing trend from 2010, but even so, it is

closer to the official rate than the labor poverty rate. The alternative SVM approach estimates the greatest increase of all four algorithms in the world financial crisis, 3.8 pp, which is still far from the labor poverty increase of 6.4 pp. The SVM ex-ante poverty predictions are the most distant of all the algorithms from their corresponding ex-post updates, with an average absolute distance of 1.36 pp. Quarterly averages of SVM estimates in panel (b) reaffirm that although the ex-post estimates generally improve the ex-ante predictions, they are behind the benchmark model and RF.

4.4 *Poverty dynamics in the COVID-19 pandemic*

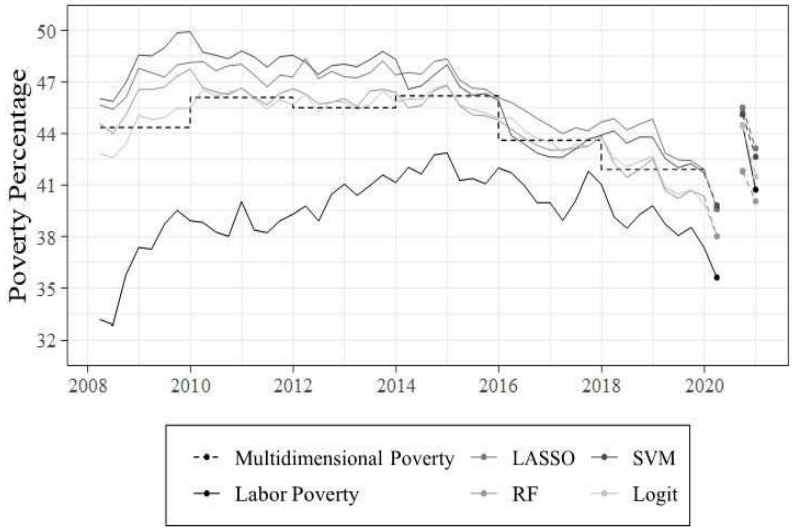
As the COVID-19 pandemic evolves, the alleviation of poverty has been threatened by disruptions in living standards and the economy.[22] This crisis is an excellent example of a situation where timely information about multidimensional poverty is fundamental to targeting policy. In order to estimate the dynamics of the pandemic, ex-ante poverty rates are provided for 2020 using the fitted models with the 2018 training data. Unfortunately, INEGI canceled the ENOE during the second quarter of 2020, but implemented a hybrid version, ENOEN, in the third quarter, 72% of whose sample is compiled from face-to-face interviews and 28% from telephone interviews.

Figure 10 shows LASSO logistic, RF, and SVM quarterly ex-ante estimates during 2020 and their corresponding ex-post poverty rates along with the labor poverty series. The pandemic seems to affect Mexican poverty after the first quarter of 2020, in line with the first national lockdown campaign in late March. Although the impact cannot be fully estimated since there is no information for the second quarter, the increase from 2020 QI to 2020 QIII is likely a lower bound of the true effect, considering that other macroeconomic variables, such as the unemployment rate and Mexican GDP, were at their worst levels in 2020 QII.[23]

[22] A recent report by the Pew Research Center estimates that there was a 19.5% increase in the number of poor people worldwide in 2020 over pre-pandemic projections (Pew Research Center, 2021). In addition, the World Bank has warned that its goal of reducing global extreme poverty to less than 3 percent by 2030 will be more difficult to reach given current challenges (World Bank, 2020).

[23] The terms "effect" and "impact" are used here as synonyms of change, not necessarily strict causality.

Figure 10
Ex-post poverty series of machine learning and 2020 ex-ante estimates



Note: Because of the COVID-19 pandemic, the ENOE was canceled in the second quarter of 2020, and replaced with a hybrid version (ENOEN) in subsequent quarters.

Source: Author's calculations using the MCS-ENIGH, ENOE, and ENOEN from the first quarter of 2008 to the last quarter of 2020.

Ex-ante numbers are presented in table 8. Labor poverty shows the highest increase from 2020 QI to 2020 QIII, approximately 8.9 pp, while the other increases range from 3.8 pp to 6.4 pp. These increases are an average of 2.5 pp larger than the corresponding impacts of the world financial crisis, though in 2020 the increasing trend is of shorter duration. The preferred projections for the 2020 multidimensional poverty rate are the average of the third and fourth quarters of 2020, with which RF predicts the lowest poverty rate (40.92%) and the LASSO logistic model the highest (44.32%). Either way, it is highly likely that the multidimensional poverty rate for 2020 will be greater than in 2018. This would also depend on further methodological changes in CONEVAL's poverty identification strategy.

Table 8
Labor poverty and ex-ante poverty estimates for 2020

2020 Quarter	*Logit*	*LASSO*	*RF*	*SVM*	*Labor Poverty*
I	38.02	39.60	38.02	39.81	35.61
III	44.43	45.49	41.79	45.12	44.46
IV	41.48	43.14	40.05	42.64	40.73
Difference QIII-QI	6.41	5.89	3.77	5.31	8.85
Mean QIII and QIV	42.96	44.32	40.92	43.88	42.60

Note: Ex-ante predictions are made with models trained with 2018 data. Because of the COVID-19 pandemic, the ENOE was canceled in the second quarter of 2020 and replaced with a hybrid survey (ENOEN) in the quarters that followed.

Source: Author's calculations using the 2018 MCS-ENIGH, ENOE, and ENOEN from the first to the last quarter of 2020.

4.5 *Model assessments*

All quarterly ex-post poverty rates seen in figure 10 are closer to the official biennial levels than the ITLP labor poverty rate, and they display similar behaviors over the period analyzed. It seems, however, that none of the ML algorithms greatly exceeds the baseline logit model in approximating the multidimensional poverty rate, so what is the real advantage of having more complex classification models? To answer this question, it would be helpful to keep in mind an observation made by Olivier Dupriez, Lead Statistician at the Development Data Group, at the 2018 Machine Learning and the Future of Poverty Prediction conference hosted by the World Bank in Washington, D.C. Dupriez presented an empirical comparison of several ML classification algorithms using data from Indonesia and Malawi, and highlighted the importance of assessing ML algorithms with multiple performance metrics.[24] Despite making good predictions of the official poverty rate, he noted, there is no guarantee that the model correctly classifies poor individuals: "The people identified as poor might be in the right numbers but they are not the right people".

[24] The conference is available at: https://www.worldbank.org/en/news/video/2018/02/27/machine-learning-future-of-poverty-prediction.

In this sense, the four models are fitted with default settings for their tuning parameters and assessed via 10-fold CV using the accuracy metric (basically 1-*Error rate*), and six complementary metrics: recall, specificity, precision, negative predictive value (NPV), F1 score, and kappa (κ). None of these metrics is better than the others; they simply evaluate different things and provide different, but equally important information.[25] Local rankings are then made for every metric and year to establish the reliability and robustness of the algorithms, and finally they are all summarized by their mean rank.

Table 9 shows these results, where RF shows the highest scores over the period analyzed, followed by SVM, the LASSO logistic model, and logit, reinforcing the findings already presented. For 2016 in particular, RF leads the rest, but the logit and LASSO overtake SVM in five metrics, possibly because of the dispersion in the 2016 training data. Interestingly, if a given model is assessed with the same metric over time, 2016 and 2018 turn out to be the years with the lowest scores, while all of them show their best performance in 2008 and 2010. The effect, however, of the 2016 methodological changes on the performance of the algorithm is more perceptible in SVM, where for instance the recall score falls from 86.7% in 2014 to 72.6% in 2016. That is, the proportion of the truly poor who are correctly classified by SVM decreases by approximately 14 pp from 2014 to 2016.

The SVM, the LASSO logistic regression, and the logit model show similar performance in the multidimensional poverty classification task in Mexico, while RF outperforms all of them. Although the logit performs well in targeting the official biennial poverty rate, the real advantage of using RF lies in its ability to correctly identify poverty patterns that are imperceptible with other algorithms, as well as its flexibility and robustness to different training data distributions and methodological changes in the poverty detection mechanisms. A state-level analysis rather than a national one, as well as consideration of the correlations between predictors, could be additional ways to explore the strengths of RF.

[25] See Appendix D for a detailed description of the performance metrics.

Table 9
Poverty classification performance

	Out-of-sample assessment metrics							
	Accuracy	*Recall*	*Specificity*	*Precision*	*NPV*	*F1 score*	κ	*Mean Rank*
				2008				
RF	0.944	0.943	0.943	0.927	0.957	0.935	0.886	1
SVM	0.890	0.884	0.884	0.860	0.913	0.872	0.775	2
LASSO	0.883	0.870	0.870	0.856	0.903	0.863	0.761	3
Logit	0.871	0.863	0.863	0.839	0.896	0.850	0.737	4
				2010				
RF	0.941	0.947	0.936	0.928	0.953	0.938	0.882	1
SVM	0.889	0.898	0.880	0.868	0.908	0.883	0.777	2
LASSO	0.880	0.882	0.877	0.863	0.895	0.873	0.758	3
Logit	0.867	0.878	0.857	0.844	0.889	0.861	0.734	4
				2012				
RF	0.938	0.948	0.929	0.922	0.952	0.935	0.875	1
SVM	0.879	0.889	0.870	0.859	0.898	0.874	0.757	2
LASSO	0.871	0.874	0.869	0.856	0.886	0.865	0.742	3
Logit	0.860	0.873	0.848	0.836	0.882	0.854	0.719	4
				2014				
RF	0.934	0.939	0.930	0.919	0.948	0.929	0.868	1
SVM	0.872	0.867	0.876	0.855	0.886	0.861	0.742	2
LASSO	0.868	0.863	0.871	0.850	0.883	0.857	0.734	3
Logit	0.856	0.863	0.850	0.829	0.880	0.846	0.711	4

Table 9
(Continued)

	Out-of-sample assessment metrics							
	Accuracy	*Recall*	*Specificity*	*Precision*	*NPV*	*F1 score*	κ	*Mean Rank*
				2016				
RF	0.939	0.935	0.941	0.917	0.954	0.926	0.874	1
LASSO	0.863	0.834	0.883	0.833	0.884	0.833	0.717	2.3
Logit	0.853	0.830	0.868	0.814	0.880	0.822	0.696	3.3
SVM	0.837	0.726	0.914	0.854	0.827	0.785	0.655	3.4
				2018				
RF	0.935	0.928	0.940	0.911	0.952	0.920	0.865	1
SVM	0.873	0.841	0.894	0.841	0.894	0.841	0.735	2
LASSO	0.866	0.830	0.890	0.833	0.887	0.832	0.720	3
Logit	0.854	0.823	0.875	0.814	0.881	0.819	0.697	4

Note: Scores and rankings of LASSO logistic regression, logit, RF, and SVM (radial basis kernel) correspond to fitted models using 50% cutoffs. Neither customized tuning grids, optimal cutoffs, nor alternative approaches for SVM are used in this comparison, and all algorithms are fitted with their default settings. The LASSO logistic specification includes interactions and quadratic terms, as in section 4.1. Training times for a single-year 10-fold CV were: SVM, 395 min.; RF, 148 min.; LASSO logistic model, 6 min.; and logit, 2 min.

Source: Author's estimates using the ENIGH-MCS from 2008 to 2018, and the caret package (version 6.0-78) in R.

5. Concluding remarks

It is currently difficult to evaluate and monitor in a timely way the progress of policies targeting poverty in Mexico using the official biennial poverty measurement reported by CONEVAL. This study shows that logistic regression and machine learning methods such as LASSO, random forest, and support vectors machine provide an innovative, efficient, and low-cost set of tools for estimating and predicting multidimensional poverty on a quarterly basis. One of the main contributions of this paper is that it shows how ex-ante estimates, which can be thought of as quarterly predictions of the subsequent level of the multidimensional poverty rate, can provide a very good estimate of poverty more than a year ahead of the official estimate. The ex-post estimates update and, in most cases, improve upon their corresponding ex-ante estimations.

Looking at out-of-sample performance, the benchmark logit model has the highest average error rate (14.6%) of the four algorithms, but its poverty estimates are closer to the multidimensional poverty rate than the rest. RF has the lowest error rate, with an average probability of misclassification equal to 0.05, surpassing the logit model by an average of 9 percentage points. In addition to their proximity to the official poverty rate, RF ex-ante and ex-post poverty estimates are also the most consistent of the four models, with an average gap of 0.5 pp. The SVM cell approach with radial basis kernel and the LASSO logistic regression present similar performance, with mean error rates of approximately 13%, and generally overestimate the multidimensional poverty levels. All of the ex-post estimates show increasing trends from the beginning of 2008 through the end of 2009, estimating an increase in Mexico of 2.3 pp to 3.8 pp in multidimensional poverty during the global financial crisis. From 2010 to 2015, most of these estimates are steady, and then they continue with downward trends until the first quarter of 2020. For the third quarter of 2020, by which time the COVID-19 pandemic had fully hit the Mexican economy, ML ex-ante poverty rates estimate an impact of 3.8 pp to 6.4 pp, while their aggregate measures forecast the 2020 multidimensional poverty rate to fall within the range of 40.9% to 44.3%.

The overall assessment of the performance of the algorithms shows that RF is the machine learning algorithm with the best measures of accuracy, recall, specificity, precision, negative predictive value, F1 score, and , followed by SVM with radial basis kernel, the LASSO logistic model, and logit, underlining the robustness of RF as a prediction method. In short, although the benchmark model does a good job of estimating the official biennial poverty rate, it is likely that

its observations labeled as poor are misclassified. It is thus preferable to use a more reliable algorithm, such as RF, if a deeper analysis of poverty profiles is required. In RF estimates, the indicator variable of per capita real labor income below the minimum welfare line, labor income, household size, state of residence, the indicator variable of rural localities, and IMSS affiliation are the most important variables for multidimensional poverty prediction, while gender, educational level, not in school, no health insurance, and economic status are the least important variables.

Future research should be directed at imputing ENOE labor income to analyze the change in results, given the growing income misreporting trend in this survey (Campos-Vázquez, 2013). The inclusion of additional models, such as conditional trees and neural networks, as well as an analysis of transition probabilities and a state-level analysis could also provide valuable information.

Acknowledgements

I thank Cecilia García, Raymundo Campos, and Eneas Caldiño, participants at the 24th Annual LACEA Meeting, and two anonymous referees for insightful comments and suggestions. All remaining errors are my own.

Ratzanyel Rincon: rrinconv@student.ubc.ca

References

Babenko, B., J. Hersh, D. Newhouse, A. Ramakrishnan, and T. Swartz. 2017. Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in Mexico, ArXiv preprint, arXiv:1711.06323.

Breiman, L. 2001. Random forests, *Machine Learning*, 45(1): 5-32.

Cámara de Diputados. 2004. Ley General de Desarrollo Social, https://www.diputados.gob.mx/LeyesBiblio/pdf/LGDS.pdf.

Campos-Vázquez, R.M. 2013. Efectos de los ingresos no reportados en el nivel y tendencia de la pobreza laboral en México, *Ensayos Revista de Economía*, 32(2): 23-54.

CONEVAL. 2010. Tendencias económicas y sociales de corto plazo y el índice de la tendencia laboral de la pobreza (ITLP), https://www.coneval.org.mx/rw/resource/coneval/med_pobreza/TendencialaboralpobrezaCONEVAL.pdf.

CONEVAL. 2014. Metodología para la medición multidimensional de la pobreza en México, https://www.coneval.org.mx/Informes/Coordinacion/Publicaciones%20oficiales/MEDICION_MULTIDIMENSIONAL_SEGUNDA_EDICION.pdf.

CONEVAL. 2019. 10 años de medición de pobreza en México, avances y retos en política social, https://www.coneval.org.mx/SalaPrensa/Comunicadosprensa/Documents/2019/COMUNICADO_10_MEDICION_POBREZA_2008_2018.pdf.

CONEVAL. 2020. Información referente al índice de tendencia laboral de la pobreza al cuarto trimestre de 2019, https://www.coneval.org.mx/SalaPrensa/Comunicadosprensa/Documents/2020/Comunicado_04_ITLP_4TRIM_2019.pdf.

Cortes, C. and V. Vapnik. 1995. Support-vector networks, *Machine Learning*, 20(3): 273-297.

Cruz, J.A. and D.S. Wishart. 2006. Applications of machine learning in cancer prediction and prognosis, *Cancer Informatics*, 2: 59-77.

EQUIDE. 2021. Resultados de la encuesta de seguimiento de los efectos del COVID-19 en el bienestar de los hogares mexicanos, http://www.equide.org/wp-content/uploads/2021/06/Comunicado_Encovid_21mar.pdf.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, 33(1): 1-22.

Guo, G., S.Z. Li, and K.L. Chan. 2001. Support vector machines for face recognition, *Image and Vision Computing*, 19(9): 631-638.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, Springer.

James, G., D. Witten, R. Tibshirani, and T. Hastie. 2013. *An Introduction to Statistical Learning with Applications in R*, New York, Springer.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. 2018. Human decisions and machine predictions, *Quarterly Journal of Economics*, 133(1): 237-293.

Lantz, B. 2015. *Machine Learning with R*, Second Edition, Birmingham, Packt Publishing.

McBride, L. and A. Nichols. 2018. Retooling poverty targeting using out-of-sample validation and machine learning, *World Bank Economic Review*, 32(3): 531-550.

Mullainathan, S. and J. Spiess. 2017. Machine learning: An applied econometric approach, *Journal of Economic Perspectives*, 31(2): 87-106.

Pew Research Center. 2021. The pandemic stalls growth in the global middle class, pushes poverty up sharply, https://www.pewresearch.org/global/2021/03/18/the-pandemic-stalls-growth-in-the-global-middle-class-pushes-poverty-up-sharply/.

Sohnesen, T.P. and N. Stender. 2016. Is random forest a superior methodology for predicting poverty? An empirical assessment, *Poverty and Public Policy*, 9(1): 118-133.

Steinwart, I. and P. Thomann. 2017. liquidSVM: A fast and versatile SVM package, ArXiv preprint, arXiv:1702.06899.

Thoplan, R. 2014. Random forests for poverty classification, *International Journal of Sciences: Basic and Applied Research*, 17(2): 252-259.

Welch, B.L. 1947. The generalisation of student's problems when several different population variances are involved, *Biometrika*, 34(1): 28-35.

World Bank. 2015. Measuring poverty https://www.worldbank.org/en/topic/measuringpoverty.
World Bank. 2020. *Poverty and Shared Prosperity 2020: Reversals of Fortune*, Washington D.C., World Bank.