Agrociencia

PREDICTION OF SOCIAL LAG IN MEXICO: A MACHINE LEARNING APPROACH FROM ECONOMIC UNIT DATA

Pablo Rodrigo **Ávila-Solís**[1], Juan Manuel **González-Camacho**[1*], Delfino **Vargas-Chanes**[2], Paulino **Peréz-Rodríguez**[1]

[1] Colegio de Postgraduados Campus Montecillo. Posgrado en Socieconomía, Estadística e Informática-Cómputo Aplicado. Carretera México-Texcoco km 36.5, Montecillo, Texcoco, State of Mexico, Mexico. C. P. 56230.

[2] Universidad Nacional Autónoma de México. Programa Universitario de Estudios del Desarrollo. Avenida Universidad, 3000, Ciudad Universitaria, Coyoacán, Mexico City, Mexico. C. P. 04510.

* Corresponding author: jmgc@colpos.mx

Citation: Rodrigo Ávila-Solís P, González-Camacho JM, Vargas-Chanes D, Peréz-Rodríguez P. 2022. Prediction of social lag in Mexico: a machine learning approach from economic unit data. Agrociencia https://doi.org/10.47163/agrociencia.v56i2.2768

Editor in Chief:
Dr. Fernando C. Gómez Merino

Received: September 01, 2021.
Approved: March 08, 2022.

Estimated publication date:
April 18, 2022.

ABSTRACT

Social lag in Mexico is officially calculated at the municipal level every five years, based on data from the population and housing census, by the National Council for the Evaluation of Social Development Policy (Consejo Nacional para la Evaluación de la Política de Desarrollo Social, CONEVAL). However, it is advisable to have annual forecasts of social lag for the follow-up of public policies. This study presents a machine learning approach to predict the classes or degrees of social lag (high, medium, low) at the municipal level in Mexico, based on information on economic units from the 2015 National Statistical Directory of Economic Units (Directorio Estadístico Nacional de Unidades Económicas, DENUE). Three supervised machine learning classifiers were implemented: logistic regression, support vector machine and random forests; and they were trained and tested in prediction based on information from counts of economic units, in their different categories, population and geographic coordinates of the municipalities; likewise, objective social lag classes were used at the municipal level, calculated with 2015 census information, reported in the literature. The criteria for evaluating the performance of the classifiers were the F-macro value, the overall accuracy and the area under the curve (ROC). The results indicate that the best overall performance was obtained with the random forest classifier with an F-macro value of 0.713 and an overall classification accuracy of 0.716; and F1-macro values for the high, medium, and low social lag classes of 0.596, 0.730, and 0.822, respectively. This classifier was used to predict social lag for 2016 and 2017. The results showed that there is a relationship between social lag and economic units aggregated at the subsector level, and that the proposed approach represents a viable and low-cost alternative for predicting social lag when census information is lacking.

Keywords: decision trees, poverty, classification models, logistic regression, artificial intelligence.

INTRODUCTION

According to the General Law of Social Development in force in 2004, federal resources and support in Mexico are distributed at the municipal level and are allocated according to the degree of social lag. It is therefore important to periodically monitor this variable in order to identify areas of high lag. Social lag, unlike poverty, is a socioeconomic indicator of social deprivation. The Mexico's National Council for the Evaluation of Social Development Politics (Consejo Nacional para la Evaluación de la Política de Desarrollo Social; CONEVAL, 2016) reported the grouping of social lag into five classes: very low, low, medium, high and very high, based on a principal component analysis, the stratification technique proposed by Dalenius and Hodges (1959) and information from the 2015 intercensal survey. This survey includes 11 sociodemographic variables related to education, health, basic services, household assets and housing quality.

The estimation of social lag is carried out every five years in an official manner, due to the availability of socioeconomic variables. Vargas-Chanes and Valdés-Cruz (2019) proposed an alternative method to estimate social lag at the municipal level, they defined three classes (low, medium and high) based on a latent class analysis and data from the 2015 intercensal survey reported by CONEVAL (2016), to give greater statistical support to the estimation of social lag at the municipal level.

Indirect methods for estimating poverty-related socioeconomic indicators are reported in the literature. Engstrom *et al.* (2017) conducted a proposal with a multiple regression model with L1 penalty to estimate the average poverty and consumption rates of 1291 municipalities in Sri Lanka from satellite imagery. These authors found that the features obtained from the images explain about 60 % of the variation in the data studied. Alsharkawi *et al.* (2021) applied 16 supervised learning models, based on different national household surveys, to classify the poverty status of households in Jordan. These include logistic regression, gradient boosting, random forests and support vector machine. These authors reported that the gradient boosting classifier obtained 81 % performance with the F1 metric with an unbalanced data set. Sani *et al. (2018)* used three classifiers to categorize monthly household income for three states in Malaysia*:* k-nearest neighbors, decision trees and empirical Bayesian. The decision tree classifier was the most significant in detecting whether households were in the poorest 40 % or outside the poorest 40 %, with unbalanced classes.

Powell *et al.* (2007) studied the availability of economic units in the United States such as grocery stores aggregated by zip code. These authors applied multivariate regression analysis to associate neighborhood location characteristics from 2000 census data, and indicated that supermarkets and grocery stores are located in low-income neighborhoods. This shows the possibility of calculating socioeconomic indicators of a geographic region from economic units.

The objective of this study was to implement three machine learning models: logistic regression, support vector machine and random forests to predict the social lag classes (low, medium and high) at the municipal level from 2015 data of economic units,

population and geographic coordinates of the municipalities; to train and evaluate in prediction with the 2015 target social lag classes reported by Vargas-Chanes and Valdés-Cruz (2019); and to perform based on the proposed approach predictions of the degrees of social lag at the municipal level for the years 2016 and 2017.

MATERIALS AND METHODS

Data collection

To conduct the study, data on economic units (UE), approximately five million records, were obtained from the fifth edition of the National Statistical Directory of Economic Units (Directorio Estadístico Nacional de Unidades Económicas, DENUE) (INEGI, 2015). INEGI has been conducting the DENUE since 2010 and as of 2015 it has been updated at least once a year. The UE were ranked according to the North American Industrial Classification System (Sistema de Clasificación Industrial para América del Norte - SCIAN). SCIAN). INEGI (2013) described this classification in five levels: sector, sub-sector, branch, sub-branch and class (Table 1). A sector consists of subsectors, a subsector is divided into branches, a branch derives into sub-branches and a sub-branch is broken down into classes. Each UE has the following attributes: unique identifier, SCIAN activity code, entity key and municipality key. The geographic location and municipal population were obtained from the national geostatistical framework (INEGI, 2014).

As target class labels, the typologies of social lag at the municipal level, low (B), medium (M) and high (A) considered in the database described by Vargas-Chanes and Valdés-Cruz (2019) were used. In the information sources, the municipality key was used and 2457 municipalities were identified; of which 971 (39.51 %) correspond to class B, 1064 (43.30 %) to class M, and 422 (17.17 %) to class A. Therefore, there were unbalanced target classes.

Database processing

Database collection, processing and integration was performed in the Python v. 3.8 programming language. The machine learning models were implemented in the Scikit-learn v. 0.23.2 platform (Pedregosa *et al*., 2011). A MacBook Air 2017 computer

Table 1. Hierarchical structure of the North American-Mexico industrial classification system to identify economic units; digits and number of categories by level of aggregation.

Digits by activity	Aggregation level	Economic categories
2	Sector	20
3	Sub-sector	94
4	Branch	303
5	Sub-branch	614
6	Class	1059

Source: INEGI (2013).

with macOS 10.14 Mojave 64-bit operating system, Intel Core i5 1.8 GHz processor and 8 GB of RAM was used to run the processes.

Using the 2015 UE data and the 2014 national geostatistical framework, for each level of aggregation and each municipality, the rate of municipal UE categories (TUE) was calculated, defined as:

$$TUE_{c,m} = \frac{UE_{c,m}}{P_m} x\ 1000$$

where $TUE_{c,m}$ is the *UE* rate for the *c*-th categroy (of the *k* aggregation level in the SCIAN) and *m*-th municipality, per thousand inhabitants; $c = 1,2,...nk$; $nk = (20, 94, 303, 614, 1059)$; $m = 1,2,..., 2457$; $UE_{c,m}$ is the UE count for specific *c* and *m*, in each level *k* and; P_m is the population of the municipality.

Five data entry scenarios were defined: sector, sub-sector, branch, sub-branch and class, according to the five hierarchies defined by SCIAN. The variables *TUE* were calculated for the economic categories defined in each scenario and two geographic variables, latitude and longitude of each municipality, were added. Thus, the first data input scenario consisted of 22 input variables (20 *TUE* + 2) for the sector level, 96 for sub-sector, 305 for branch, 616 for sub-branch; and 1061 for class.

Machine learning classifiers

Three machine learning models were implemented: logistic regression (LR); support vector machine (SVM); and random forests (RF) for the different input scenarios.

Logistic regression model

The linear binary classifier LR is based on the statistical theory of generalized linear models and uses the link function *logit* (McCullagh y Nelder, 1989). Which is defined as:

$$logit(p) = log\left(\frac{p}{1 - p}\right)$$

where *p* is the probability of the event of interest.

Then, LR is modeled as the linear combination of the parameters and the input characteristics, i.e:

$$logit(p^{(i)}) = \sum_{j=0}^{m} w_j x_j^{(i)} = z^{(i)}$$

where w_j is the *j*-th parameter and $x_j^{(i)}$ is the *j*-th characteristic of the *i*-th sample of the training data set.

In this study, the LR model was considered with an L2 regularization term (James *et al*., 2019) to avoid overfitting the model to the training data; which consists of optimizing the loss function $J(w)_{LR}$ expressed by:

$$J(w)_{LR} = \sum_{i=1}^{n}\left[-y^{(i)}\log\left(\phi\left(z^{(i)}\right)\right) - \left(1-y^{(i)}\right)\log\left(1-\phi\left(z^{(i)}\right)\right)\right] + \frac{\lambda}{2}\sum_{j=1}^{m} w_j^2$$

where $\phi(z) = \frac{1}{1+e^{-z}}$ is the sigmoid activation function; w_j the *j*-th parameter, note that w_0 it is not penalized; $y^{(i)}$ is the target class of the *i*-th sample, which considers the values 0 and 1; λ is the regularization hyperparameter L2; *n* is the number of samples or observations; and *m* is the number of parameters. This model is generalized for multiple classes, by means of the multiple classification approaches: one versus the rest and one versus one, combining binary classifiers.

Another approach to multi-class classification with this model is addressed by means of the *softmax* function, which is defined as,

$$p\left(y=k \mid z\right) = \phi(z) = \frac{e^{z}k}{\sum_{j=1}^{K} e^{z_j}}$$

where *K* represents the number of target classes; *k* is a particular class to calculate the softmax function; and z_k is the linear combination of a particular class with its own vector of weights. *Softmax* obtains the belonging probabilities for each class and the maximum probability is used to make a prediction. The cost function used for RL is the cross-entropy function, which is defined as:

$$j(w) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log\left(p\left(y=k \mid z\right)\right)$$

where $y_k^{(i)}$ is the probability that the sample *i* belongs to the class *k*; *m* is the number of samples; and *K* the number of target classes. In particular, if *K*=2, the cost function of the binary RL model is obtained.

Support Vector Machine

The SVM classifier consists of maximizing the margin of the hyperplane separating two classes. The margin is defined as the distance between the separation hyperplane and the training samples closest to it. SVM was used under the multiclass one versus the rest classification approach for the three classes of social lag considered in this

study. SVM with L2 regularization for margin-flexible classification, consisted of minimizing the loss function $J(w)_{SVM}$ (Wang *et al*., 2006), which is expressed by:

$$J(w)_{SVM} = \frac{1}{2} \| w \|^2 + \frac{c}{2} \sum_{i=1}^{m} \xi^{(i)^2}$$

$$t^{(i)} \left(w^T x^{(i)} + b \right) \geq 1 - \xi^{(i)}$$

where w is the vector of weights or parameters; b is the bias term; $t^{(i)}$ is the target class of the i-th sample, which considers the values -1 and 1; C is a hyperparamter that controls the width of the hyperplane margin and, hence, of the L2 regularization term; and ξ is a slack variable introduced by Cortes and Vapnik (1995) to deal with nonlinearly separable data.

In a binary classification we have two cases $w_0 + w^T x^{(i)} \geq 1 - \xi^{(i)}$ for $t^{(i)} = 1$ and $w_0 + w^T x^{(i)} \leq 1 - \xi^{(i)}$ for $t^{(i)} = -1$, where $x^{(i)}$ is the i-th sample of the training data set.
The use of kernel functions (Bishop, 2006) facilitates the solution of nonlinear classification problems by means of linear combinations of the input variables or characteristics projected into a higher dimensional space. The kernels used are linear, $K(x^{(i)}, x^{(j)}) = x^{(i)^T} x^{(j)}$; radial basis function (rbf), $K(x^{(i)}, x^{(j)}) = e^{-\gamma \| x^{(i)} - x^{(j)} \|^2}$; and sigmoid, $K(x^{(i)}, x^{(j)}) = \tanh\left(\gamma \cdot x^{(i)^T} x^{(j)} + r \right)$; where γ and r are hyperparameters that are optimized by means of a grid with intervals of discrete values.

Random forests

The RF classifier proposed by Breiman (2001) is an ensemble method that uses combinations of a given number of decision trees. These are taken from a random sample with replacement and trained on a given number of input characteristics taken from a sample without replacement. After the decision trees are generated, the majority vote is used to predict a class. This approach seeks to minimize the entropy or Gini impurity of the model's decision trees. The entropy measure $I_E(q)$ for non-empty classes is defined as,

$$I_E(q) = -\sum_{k=1}^{c} p(k \mid q) \log_2 \left(p(k \mid q) \right)$$

where $p(k \mid q)$ is the proportion of observations of a node q, which belong to the class k, such that $I_E(q) = 0$, if all examples of the node belong to the same class; and $I_E(q) = 1$ if the observations are distributed in the classes uniformly.

The Gini impurity criterion $I_G(q)$, is defined as:

$$I_G(q) = -\sum_{k=1}^{K} p(k \mid q) \left(1 - p(k \mid q) \right) = 1 - \sum_{k=1}^{K} p(k \mid q)^2$$

where $I_G(q)$ reaches its maximum value when the samples have a uniform distribution within classes (Raschka and Mirjalili, 2019).

The hyperparameters of the RF model that are optimized by means of a grid search are the number of estimators (decision trees) of the model (*NE*); the quality criterion for splitting the nodes (*CR*); the maximum depth of the decision trees (*MP*) controlling the complexity of the model; and the maximum number of characteristics or input variables (*MC*) that are randomly chosen to split the nodes.

Performance metrics

The following metrics were used to evaluate the performance of the classifiers: global accuracy (*ACC*, accuracy), precision (*P*), sensitivity (*S*), F1 score, F1-macro value, and area under the curve (*AUC*) of the ROC curve and the P-S curve.

The confusion matrix describes the class predictions (represented by columns) versus the true classes (represented by rows). For the binary case (Table 2) this matrix reports the number of true positives (*VP*), positive observations, predicted as positive; the number of false positives (*FP*), negative observations predicted as positive; the number of false negatives (*FN*), positive observations predicted as negative; and the number of true negatives (*VN*), negative observations predicted as negative.

ACC defined as:

$$ACC = \frac{VP + VN}{VP + VN + FP + FN}$$

P is the fraction of *VP* with respect to the total predicted positives and is calculated by:

$$P = \frac{VP}{VP + FP}$$

S is the fraction of *VP* with respect to the total observed positives and is defined as:

Table 2. Description of a confusion matrix for the case of a binary classifier.

		Predicted classes	
		Positive	Negative
True classes	Positive	*VP*	*FN*
	Negative	*FP*	*VN*

VP: true positive; *FN*: false negative; *FP*: false positive; *VN*: true negative.

$$S = \frac{VP}{VP + FN}$$

F1 represents the harmonic mean of *P* and *S*, that is:

$$F1 = \frac{2P\ S}{P + S}$$

where F1 takes values between 0 and 1; 0 indicates a poor fit and 1 a perfect fit. The metric F1 favors classifiers with *P* and *S* the like (Géron, 2019).

F1-*macro*, in the case of a multi-class problem, it is calculated as the arithmetic average of the measurements F1 obtained for each class:

$$F1 - macro = \frac{1}{N}\sum_{i}^{N} F1_i$$

where $F1_i$ is the value of F1 of the *i*-th target class; and *N* is the total number of classes. This metric is appropriate when there are unbalanced classes; each class has equal weight and emphasis is placed on the classes with lower frequency (Lipton *et al.*, 2014).

The ROC curve associates the rate of *VP* on the vertical axis versus the rate of *FP* on the horizontal axis. To compare the performance of the classifiers *AUC* is used, which varies between 0 and 1, values close to 1 indicate a good performance of the classifier (Fawcett, 2006). In the case of a binary classification there is one ROC curve per model, and in the multiclass case there is one ROC curve for each class of the model. The P-S curve represents the plot of values of *S* on the horizontal axis versus values of *P* on the vertical axis. The value *AUC* of the P-S curve takes values between 0 and 1, values close to 1 indicate a good performance of the model. In the binary case, when the data are not balanced, the P-S curve is more appropriate to compare the performance of the classifiers (Saito and Rehmsmeier, 2017).

Classifier training

Variable scaling

The input dataset (or characteristics) was standardized before performing the training of the LR, SVM and RF classifiers, based on the expression,

$$x_j' = \frac{x_j - \mu_j}{\sigma_j}$$

where x_j' is a vector of observations of the *j*-th input characteristic, from the training set of *n* observations; μ_j is the sample mean of x_j'; and σ_j is the sample standard deviation.

Variable scaling improves the performance of optimization algorithms. The scaling parameters μ_j and σ_j of each characteristic j are then applied to transform the test set characteristics and evaluate the performance of the classifiers.

Hyperparameter selection

The selection of optimal hyperparameters for each classifier was performed by means of a grid search and cross-validation (CV) with k=10 partitions. A random partition was used with 1965 observations (80 %) from the input data set. For each classifier, the hyperparameter combination that obtained the maximum average value of the F1-macro criterion was selected from the possible combinations in each hyperparameter grid.

The hyperparameters and value ranges analyzed for each classifier are described in Table 3. In particular, LR was trained with the inverse of the L2 regularization hyperparameter, with two classification strategies: one versus the rest and multinomial; and optimized with the L-BFGS-B algorithm (Morales and Nocedal, 2011).

Evaluation of model performance

The evaluation of the final performance of the LR, SVM and RF classifiers was performed with the complete datasets (2457 observations) and with the set of input variables or characteristics retained. The CV procedure was performed with k-5 random partitions, stratified by the three social lag classes to obtain the average performance metrics. The predictive capacity of the three classifiers was evaluated with the same input scenarios (categories in the SCIAN) and random partitions of the data to avoid biases in the comparison of their predictive performance.

Table 3. Hyperparameters and value ranges, from logistic regression (LR), support vector machine (SVM) and random forest (RF) classifiers.

Model	Hyperparameter	Range
LR	λ	0.001, 0.01, 0.1, 1, 10, 100, 1000 and 10 000
	method	one *versus* the rest and multinomial
SVM	Kernels	Linear, sigmoid and rbf
	C	0.0001, 0.001, 0.01, 0.1, 1, 10, 100 y 1000
	γ	0.0001, 0.001, 0.01, 0.1, 1, 10, 100 y 1000 (kernels sigmoid and rbf)
	r	0, 1, 2,3, 4, y 5 (kernel sigmoid)
RF	*NE*	100, 150, 200, 250, 300, 350, y 400
	CR	Entropy and Gini
	MC	Square root and logarithm base 2
	MP	10, 15, 20, 25, y 30

NE: number of estimators; *CR*: division criterion; *MC*: maximum number of characteristics; *MP*: maximum depth. The ranges of values were selected heuristically and experimentally.

RESULTS AND DISCUSSION

The optimized data input scenarios and hyperparameters of the classifiers were as follows: LR (at subsector level, λ-10); SVM (at subsector level, sigmoid kernel, C=100, r = 0 and γ = 0.0001); and RF (at subsector level, entropy *CR*, *MP* = 20, $MC = \sqrt{96}$ and *NE* = 250). According to the optimized L2 regularization hyperparameters, LR required more regularization than SVM. Of the five data input scenarios analyzed, the SCIAN subsector aggregation level obtained the best predictive performance in all three models.

RF outperformed LR and SVM with the metrics F1-*macro* and *ACC*. RF outperformed LR and SVM for *P* in classes B and A; for *S* in class M; and for F1 in classes B and M. LR outperformed SVM and RF for *P* in class M, and for *S* in class B. SVM outperformed LR and RF for *S* and F1 in class A. Furthermore, it was observed that all three classification models performed better in identifying class B (Table 4). For the social lag class A, RF minimized *FP* (*P*=0668); that is, RF obtained the best performance for predicting class A. In turn, SVM minimized *FN* (*S* = 0.616); that is, SVM identified the largest number of true A classes.

The ROC curves for each target class of the RF classifier show that the maximum value of *AUC* corresponded to class B and the minimum value corresponded to class M; while the P-S curves show that RF obtained better performance for classifying class B and the lowest performance was obtained with class A (Figure 1). The optimistic view of the ROC curve is attributed to the non-uniform distribution of the three classes of social lag to evaluate the performance of the classifiers. In this type of situation there are more robust alternatives such as the concentrated ROC curve, the cost curve and the P-S curve, the latter option is considered the most appropriate (Saito and Rehmsmeier, 2015). Another way to measure the performance of the RF classifier is by means of the normalized confusion matrix, which describes the proportion of true classes within each predicted class; this matrix shows the percentages of observations that belong to the correctly predicted class; these are 78, 79, and 54 %, for classes B,

Table 4. Average prediction performance criteria of logistic regression (LR), support vector machine (SVM) and random forest (RF) classifiers with cross-validation (*k*=5).

Model	*F*1-*macro*	*ACC*	Class	*P*	*S*	F1
			B	0.839	0.786	0.812
LR	0.709 +/- 0.028	0.732 +/- 0.016	M	0.687	0.737	0.711
			A	0.617	0.595	0.606
			B	0.836	0.777	0.805
SVM	0.708 +/- 0.024	0.728 +/- 0.016	M	0.685	0.727	0.706
			A	0.612	0.616	0.614
			B	0.864	0.785	0.822
RF	0.716 +/- 0.021	0.744 +/- 0.010	M	0.679	0.789	0.730
			A	0.668	0.538	0.596

ACC:global accuracy; target classes of social lag (low: B, medium: M, and high: A); precision (*P*); sensitivity (*S*); and metric *F*1.

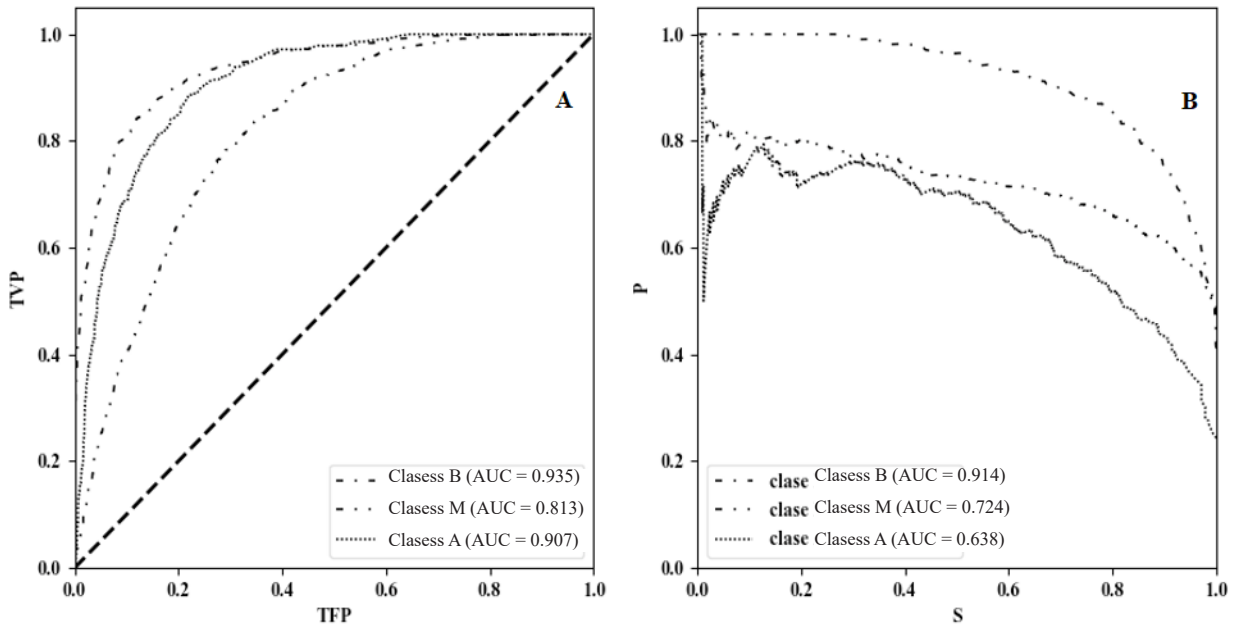

Figure 1. Performance curves of the random forest (RF) classifier for predicting social lag classes in Mexico (low: B, medium: M, and high: A) in Mexico. A: ROC curve and B: P-S curve; P: precision; S: sensitivity; TVP: true positive rate; TFP: false positive rate; AUC: areas under the ROC curve; and P-S curve.

M, and A, respectively (Figure 2). Classes M and B had better prediction than class A; samples from class M were mistaken for class A.

The predictions of the social lag classes of the RF model from EU aggregated at the subsector level were used to make a comparison of the geographic distribution at the municipal level with the lag classes reported by Vargas-Chanes and Valdés-Cruz (2019). The maps show that the geographic distribution of class B is mostly located in northern Mexico and in both maps the distributions are very similar. The distribution of class M is very similar in the center and southeast, and in the north in some municipalities, the RF model classifies them as class B or class A. The distribution of class A is very similar in central and southeastern Mexico; however, in the north, RF classifies some class A municipalities as class M (Figure 3).

The RF model optimized with 2015 data was used to make the predictions of the degrees of social lag from UE of October 2016 and November 2017 data from the DENUE (Figure 4); where it is observed that the geographic distribution of the classes of social lag is very similar to that reported by Vargas-Chanes and Valdés-Cruz (2019). The application of the machine learning approach proposed in this research to predict social lag in Mexico is a novel, feasible, and low-cost approach; in particular, because of the type of data or input variables used in this work. The results show that from 2015 to 2016, there was a decrease in the proportion of municipalities with high social lag from 17.2 to 15.5 % (Table 5). That is, 42 municipalities went from a high to a medium

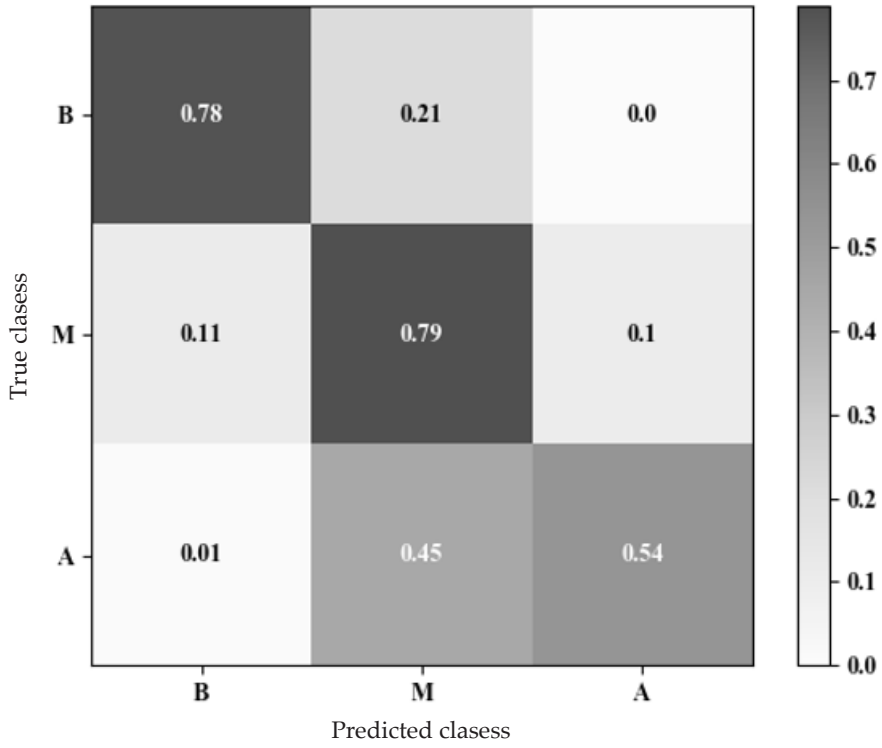

Figure 2. Normalized confusion matrix of the classifier random forests (RF) to predict social lag classes (low: B, medium: M, and high: A) in Mexico.

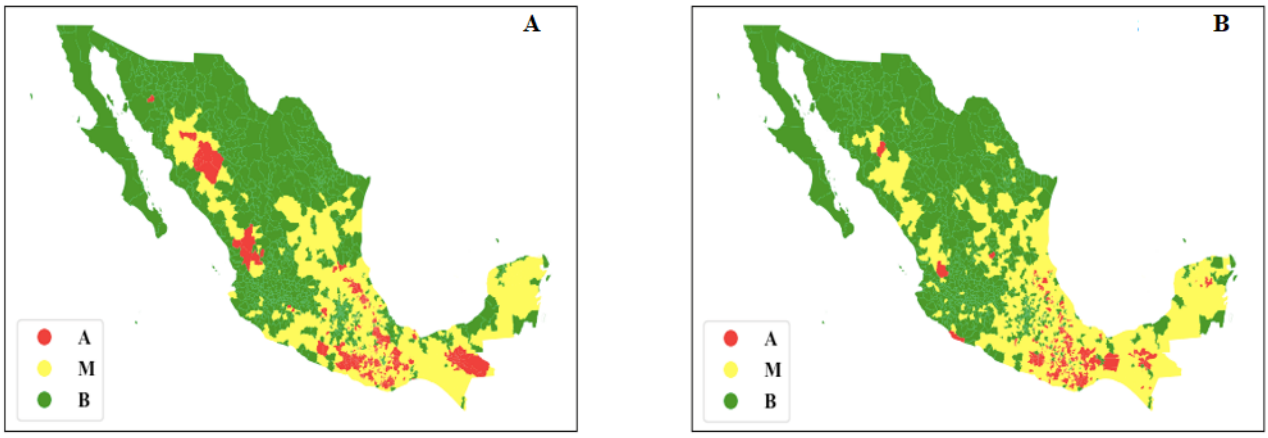

Figure 3. Municipal geographic distribution of the classes of social lag (low, medium and high): B, medium: M, and high: A) in 2015 in Mexico. A: values reported by Vargas-Chanes and Valdés-Cruz (2019) *versus* B: predictions made with the random forests (RF) classifier from 2015 data of economic units aggregated at the subsector level.

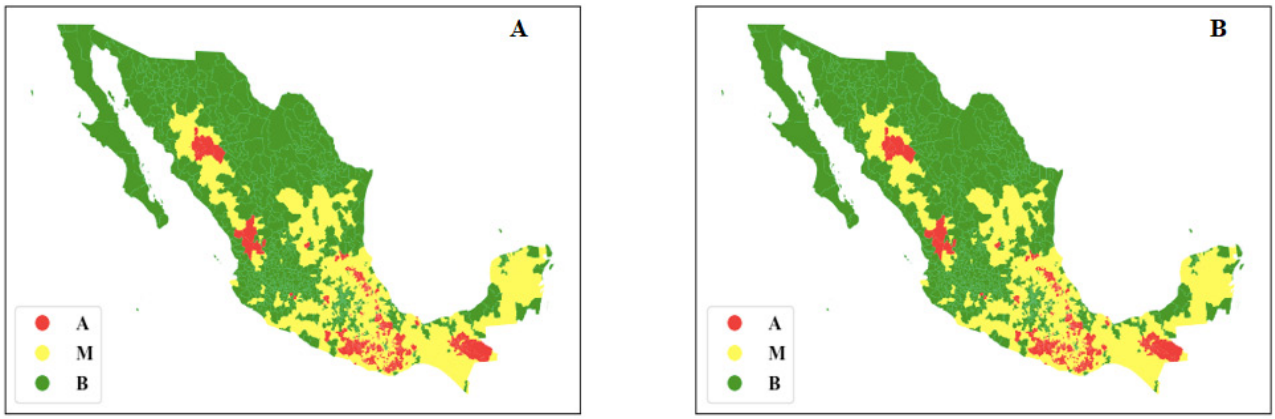

Figure 4. Municipal geographic distribution of the classes of social lag (low, medium and high): B, medium: M, and high: A) predicted with the random forest classifier (RF) in Mexico. A: from October 2016 economic unit (UE) data; and B: UE of November 2017, aggregated at the subsector level based on data from the DENUE.

Table 5. Total number of municipalities (percentages) predicted with the random forest (RF) model by category of social lag (low B, medium: M, and high: A) in 2016 and 2017 in Mexico. Year 2015 shows the RF model test data.

Year	Total and percentage of municipalities per class of social lag		
	B	M	A
2015	971 (39.5 %)	1064 (43.3 %)	422 (17.2 %)
2016	971 (39.5 %)	1106 (45.0 %)	380 (15.5 %)
2017	976 (39.7 %)	1101 (44.8 %)	380 (15.5 %)

degree of social lag, and no changes were observed in municipalities with low social lag. On the other hand, predictions for 2016 and 2017 show that municipalities with high social lag did not change, and five municipalities with medium social lag in 2016 moved to a low degree of social lag. Most of the municipalities did not present changes in the predicted degree of social lag, despite the fact that data from two consecutive years of the DENUE were used as input variables for the RF model.

The municipalities that changed their level of social lag positively from 2016 to 2017, according to the RF model are: Eduardo Neri, Guerrero; Comonfort, Guanajuato; El Marqués, Querétaro; Tapalpa, Jalisco; and Perote, Veracruz. The municipalities Eduardo Neri, Guerrero; Tapalpa, Jalisco; and Comonfort, Guanajuato had slight increases in UE in 2017, mainly in those UE related to retail trade. The municipality of El Marqués, Querétaro, significantly increased the registration of UE related to construction and industry; this was detected by the model and predicted the change. Finally, the municipality of Perote, Veracruz, presented mixed changes in the different degrees of social lag at the subsector level, so further analysis is required to explain the changes in the prediction of social lag.

Based on the above, the proposed methodology for predicting the degree of social lag based on UE information obtained results similar to traditional methodologies that use population and housing census data, which highlight the little variation in the levels of lag at the municipal level during the time period analyzed. In this context, the capacity of decision-makers responsible for mitigating social lag in Mexico is limited, particularly in areas of high social lag.

The ordinal nature of the degrees of social lag was not considered in this study in order to optimize the performance of the machine learning classifiers. As shown in other studies (Pérez-Rodríguez *et al*., 2020; Cao *et al*., 2020) the performance of classifiers improves, under the assumption of an equidistant hierarchy between the target classes under investigation. Likewise, it is possible to improve the performance of machine learning classifiers to predict a specific class or social lag category of interest, this allows determining an optimal metric for a specific class (e.g., high social lag) ; in this case, the AUC metric of the P-S curve for class A is used as a performance criterion, instead of the metric ACC used in this study.

CONCLUSIONS

The machine learning classifiers of logistic regression, support vector machine and random forests implemented in this research obtained a good prediction performance to predict the classes or degrees of low, medium and high social lag at the municipal level based on data from the national directory of economic units aggregated at the subsector level (94 subsectors) and their geographic location.

In particular, the random forest classifier obtained the best average prediction performance with an overall classification accuracy of 74.4 % and F1-macro of 71.6 %. The low and medium social lag classes were predicted more accurately than the high social lag class. The results of the study show a strong correlation between the degrees of social lag at the municipality level and the data of the corresponding economic units.

The geographic distribution of social lag classes predicted in 2015 with the random forest model from economic units was very similar to the geographic distribution of social lag reported in 2015. The predictions of social lag for 2016 and 2017 made with the proposed approach show that it is possible to predict the degree of social lag indirectly and at low cost, when population and housing census statistics are not available to determine social lag.

REFERENCES

Alsharkawi A, Al-Fetyani M, Dawas M, Saadeh H, Alyaman M. 2021. Poverty classification using machine learning: the case of Jordan. Sustainability. 13: 1–16. https://doi.org/10.3390/su13031412

Bishop CM. 2006. Pattern Recognition and Machine Learning. Springer. New York, NY, USA. pp: 291–294.

Breiman L. 2001. Random Forests. Machine Learning 45: 5–32.

Cao W, Mirjalili V, Raschka S. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters 140: 325–331. https://doi.org/10.48550/arXiv.1901.07884

CONEVAL (Consejo Nacional para la Evaluación de la Política de Desarrollo Social). 2016. Índice de Rezago Social 2015. Presentación de Resultados. https://www.coneval.org.mx/Medicion/Documents/Indice_Rezago_Social_2015/Nota_Rezago_Social_2015_vf.pdf (Retrieved: January 2021).

Cortes C, Vapnik V. 1995. Support-vector networks. Machine Learning 20: 273–297.

Dalenius T, Hodges J. 1959. Minimum variance stratification. Journal of the American Statistical Association 54: 88–101.

Engstrom R, Hersh J, Newhouse D. 2017. Poverty from Space: Using High-Resolution Satellite Imagery for Estimating Economic Well-Being. World Bank Policy Research Working Paper No. 8284. https://ssrn.com/abstract=3090770 (Retrieved: February 2021).

Fawcett T. 2006 An introduction to ROC analysis. Pattern Recognition Letters 27: 861–874.

Géron A. 2019. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems 2nd ed. O'Reilly Media. Sebastopol, CA, USA. pp: 60–66.

INEGI (Instituto Nacional de Estadística y Geografía). 2013. Sistema de Clasificación. Industrial de América del Norte, México SCIAN 2013. https://www.inegi.org.mx/contenidos/productos//prod_serv/contenidos/espanol/bvinegi/productos/clasificadores/SCIAN/SCIAN_2013/702825051693.pdf (Retrieved: January 2021).

INEGI (Instituto Nacional de Estadística y Geografía). 2014. Marco geoestadístico 2014 versión 6.2 (DENUE). https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825004386 (Retrieved: January 2021).

INEGI (Instituto Nacional de Estadística y Geografía). 2015. Directorio Nacional de Unidades Económicas DENUE. https://www.inegi.org.mx/app/mapa/denue/default.aspx (Retrieved: January 2021).

James G, Witten D, Hastie T, Tibshirani R. 2013. An Introduction to Statistical Learning: with Applications in R. Springer. New York, NY, USA. pp: 215–219.

Lipton ZC, Elkan C, Naryanaswamy B. 2014. Optimal Thresholding of Classifiers to Maximize F1 Measure. *In*: Machine Learning and Knowledge Discovery in Databases. Calders T, Esposito F, Hüllermeier E, Meo R. (eds.) ECML PKDD 2014. Lecture Notes in Computer Science 8725. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-44851-9_15

McCullagh P, Nelder JA. 1989. Generalized Linear Models, volume 37 of Monographs on Statistics and Applied Probability. 2a Ed. Chapman and Hall. London, UK. 532 p.

Morales JL, Nocedal J. 2011. Remark on "algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization". ACM Transactions on Mathematical Software 38(1): 1–4. https://doi.org/10.1145/2049662.2049669

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 2011. Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12(85): 2825–2830.

Pérez-Rodríguez P, Flores-Galarza S, Vaquera-Huerta H, Del Valle-Paniagua DH, Montesinos-López OA, Crossa J. 2020. Genome-based prediction of Bayesian linear and non-linear regression models for ordinal data. Plant Genome 13: 1–13. https://doi.org/10.1002/tpg2.20021

Powell LM, Slater S, Mirtcheva D, Bao Y, Chaloupka FJ. 2007. Food store availability and neighborhood characteristics in the United States. Preventive Medicine 44: 189–195.

Raschka S, Mirjalili V. 2019. Python Machine Learning 3rd ed. Packt Publishing Ltd. Birmingham, UK. pp: 90–96.

Saito T, Rehmsmeier M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10(3). https://doi.org/10.1371/journal.pone.0118432

Saito T, Rehmsmeier M. 2017. Precrec: fast and accurate precision–recall and ROC curve calculations in R. Bioinformatics 33: 145–147. https://doi.org/10.1093/bioinformatics/btw570

Sani NS, Rahman MA, Bakar AA, Sahran S, Sarim HM. 2018. Machine learning approach for bottom 40 percent households (B40) poverty classification. International Journal on Advanced Science, Engineering, and Information Technology 8: 1698–1705. http://dx.doi.org/10.18517/ijaseit.8.4-2.6829

Vargas-Chanes D, Valdés-Cruz S. 2019. A longitudinal study of social lag: regional inequalities of growth in Mexico 2000 to 2015. Journal of Chinese Sociology 6: 1–18. https://doi.org/10.1186/s40711-019-0100-6

Wang L, Zhu J, Zou H. 2006. The doubly regularized support vector machine. Statistica Sinica 16: 589–615.