



# Bayesian Networks and Machine Learning Approaches Applied to Social Backwardness

Jesús Alejandro Navarro-Acosta<sup>1</sup> · Jesús-Adolfo Mejía-de-Dios<sup>1</sup> · José María González Lara<sup>2</sup> · Edgar J. Sanchez Carrera<sup>1,3</sup>

Received: 17 February 2024 / Accepted: 25 September 2025  
© The Author(s) 2025

## Abstract

This paper applies Bayesian and machine learning techniques to analyze Mexico's Social Backwardness Index data from 2000 to 2020. This index aggregates key socioeconomic factors such as education, access to health services, essential housing services, housing quality and spaces, and household assets. We aim to identify the insights, such as conditional dependencies between these variables, and determine which factors most significantly contribute to social backwardness in Mexico. Through machine learning and non-parametric techniques (such as XGBoost, Neural Network Implementations, and Permutation Feature Importance), we identify which socioeconomic indicators most impact the degree of social backwardness. The Bayesian network is then employed to visualize the relationships between those socioeconomic indicators and the social backwardness index, providing information on the dependencies and linkages between features such as illiteracy, household appliances, and essential housing services. The analysis shows that critical indicators such as lack of household appliances, illiteracy, and inadequate housing services (e.g., lack of toilets and drainage) are highly predictive of social backwardness. Over the years, the importance of these variables shifts, but they remain consistently relevant in determining the level of social backwardness. Bayesian learning results suggest that policies targeting improvements in these primary household conditions could substantially reduce social backwardness across Mexico.

**Keywords** Bayesian probability · Dependencies and effects · Machine learning · Social backwardness in Mexico

---

Jesús Alejandro Navarro-Acosta, Jesús-Adolfo Mejía-de-Dios, José María González Lara, Edgar J. Sanchez Carrera Both are equally contributed to this work.

---

Extended author information available on the last page of the article

## 1 Introduction

Machine Learning (ML) is a part of artificial intelligence that aims to build models based on sample data to make predictions or decisions without being explicitly programmed. So, the system learns from its data experience. There are two main ways these models learn from the data. One of them is known as supervised learning. In supervised learning, the model receives labeled training data. The model provides previously cataloged examples, with classification and regression being the two main problems addressed in this type of learning. On the other hand, there is unsupervised learning, where the algorithm must try to interpret the data since the training data is not labeled.<sup>1</sup> For example, (Ciaburro, 2022) studies the different methodologies for identifying the most common mechanical failures. They concluded that the most widely applied algorithms based on machine learning are Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms. Zhong et al. (2021) provide a structured overview of how machine learning models can be applied to solve complex science and engineering problems. Then, they give a taxonomy of these existing techniques, which uncovers knowledge gaps and potential crossovers of methods between disciplines.<sup>2</sup>

Using ML models in applied economics is a relatively recent practice, and ML techniques have attracted increasing attention over the last decade due to their power to perform out-of-sample forecasts and discover potentially very complex structures for data whose functional form was not prespecified in socioeconomic aspects (Lazebnik, 2024; Polyzos & Siriopoulos, 2023). In recent years, moreover, ML methods have substantially improved their forecasting capacity by implementing new algorithms and have acquired a new centrality in the economic literature (Einav & Levin, 2014; Gogas & Papadimitriou, 2021; Athey, 2018; Varian, 2014; Monroe et al., 2015; Shami & Lazebnik, 2024). One of the research fields that has benefited from the arrival of machine learning algorithms and systems is the field of economic development (Hassan et al., 2024; Ozden & Guleryuz, 2022). It is noteworthy that big data from economies characterized by social pathologies of development, such as social poverty or social backwardness, allow us to diagnose and treat such social pathologies through machine learning techniques in such developing economies (Felix et al., 2024). A task of exceptional progress is the algorithms to predict diagnoses or treatments based on poverty traps (e.g., environmental traps, see Spandagos et al.

---

<sup>1</sup> Nowadays, machine learning is applied in various areas with different aims. For instance, fault detection in the industrial field is crucial because its complex processes involve the optimal operation of several machines. In (Liu et al., 2020), machine learning performs a model for wind turbine fault detection because wind turbine control aims to maintain a safe operational status while achieving cost-effective operation. For fault detection in induction machines, (Gonzalez-Jimenez et al., 2021) proposes implementing several machine learning algorithms such as logistic regression, support vector machines, random forest, and k-nearest neighbors.

<sup>2</sup> In medical and health areas, machine learning has also been used to help analyze various scenarios such as COVID disease. Kwekha-Rashid et al. (2021) conclude that machine learning can play an important role in COVID-19 research, prediction, and discrimination. Moreover, ML can also participate in health providers' programs and plans to evaluate and classify COVID-19 cases. In Chugh et al. (2021), a survey on ML in breast cancer diagnosis is presented and manifested SVM to be the finest classifier for diagnosing breast carcinoma.

(2023)). In Thalari et al. (2023), a predictive modeling of socioeconomic trends using ML is presented to analyze the possible implications of these trends in policy planning.<sup>3</sup> Radovanovic and Haas (2023) propose to integrate ML models to improve the prediction upon existing bankruptcy prediction models.<sup>4</sup> Feldmeyer et al. (2020) generate socioeconomic indicators using geodata from OpenStreetMap (OSM) and machine learning.<sup>5</sup> On the other hand, in Balasankar et al. (2020), an Intelligent socioeconomic status (SES) prediction system using ML models is presented. The aim is to understand the Socioeconomic System issues and predict SES levels in a particular region.<sup>6</sup> Storm et al. (2019) reviews the ML approaches from an applied economist's perspective, identifying current limitations of the econometric models and exploring potential solutions.

Therefore, the main goal of this research is to apply ML techniques and Bayesian network approaches to address the lack of literature on the analysis of the dependence effects of variables on the social backwardness of Mexican households. Since emerging (or developing) countries such as Mexico, where poverty traps are shown ((Brida et al., 2021; Risso et al., 2013; Sanchez Carrera & Risso, 2023)) and where social backwardness and inequality of opportunities have increased considerably in recent decades (Agüero & Beleche, 2013; Tagliati, 2022), are exciting cases to analyze in depth.

In this paper, the aim is not to compare the accuracy of the predictions obtained with various ML models concerning standard classification techniques, a fact that is widely documented, even not unanimously, in the literature, but to produce possible

---

<sup>3</sup>The authors state that although traditional economic benchmarks like Gross Domestic Product (GDP) and unemployment rates are fundamental economic well-being indicators, these often furnish a retrospective viewpoint, lacking the timeliness and granularity required for effective policy planning in a changing world. Therefore, factors such as insights from geospatial data (land usage patterns, urban development, and environmental shifts) and sentiment analysis were incorporated, i.e., a Multi-Source Data Integration was performed and then analyzed by ML models such as Gradient Boosting and Long Short-Term Memory (LSTM). The results show that using a multi-source data scheme achieves better results in predicting economic indicators (such as GDP) than using only economic indicators.

<sup>4</sup>Traditional bankruptcy prediction models typically focus solely on predicting the event of bankruptcy itself and do not consider the socioeconomic consequences of their prediction. Moreover, two alternative evaluation metrics are considered, one being the social impact measured using the number of lost jobs. After comparing several ML models (such as SVM, Linear Discriminant Analysis, Logistic Regression, etc.), the authors conclude that small differences in statistical performance can translate into large differences regarding socioeconomic costs.

<sup>5</sup>OSM is a free, editable map of the whole world that is being built by volunteers largely from scratch and released with an open-content license. The hypothesis in this research is that there are proxies for socioeconomic attributes within the geodata of the OSM database. For instance, Can the size of industrial areas or the density of public transportation or infrastructure indicate unemployment rates? The analysis used four indicators: residents, unemployment, migration, and elderly. These were selected because the official statistical data are available; therefore, it is possible to test the suitability of OSM as a data source. The authors conclude that OSM documents the physical manifestation of human activities, and these data can be used to perform socioeconomic analyses using machine learning.

<sup>6</sup>The study was carried out in Rajahmundry, AP, India, and 48 features such as family size, married people, child work below 15, annual income, etc. were considered. And one target column data with four class attributes (poor, rich, middle, upper-middle), i.e., this is a multi-class problem. Algorithms such as Naive Bayes, SVM, KNN, and RF were compared, with the RF approach achieving better scores for classification metrics.

Bayesian estimates on the probability that explains the social backwardness in socioeconomic indicators at the household level. Hence, we aim to face the research question of which variable affects and predicts the social backwardness of households, and so to offer economic policy recommendations (close to those done by Shah (1991); Sharma and Paramati (2018); Tsagris (2021); Zhang et al. (2022)). Thus, our approach is novel in applying Bayesian ML techniques to socioeconomic data in Mexico. In this vein, we want to handle various potential types of non-linearities in the data that standard statistical inference techniques find challenging to do without making any assumptions and being limited to assuming a functional form for these non-linearities, as happens when one uses a standard regression technique.

To achieve our goal, we consider the database of the National Council for the Evaluation of Social Development Policy (CONEVAL) in Mexico (CONEVAL, 2023), as it constructs an Index of Social Backwardness (SBI) that summarizes aggregate indicators of access to some of the social rights of people and their well-being at home for the different geographical disaggregations (federal entities, municipalities, and localities). This allows monitoring of four relevant indicators: i) educational lag, ii) access to health services, iii) the quality and spaces of the home, and iv) essential services in housing. In addition, indicators referring to household assets are in the analysis. In addition, CONEVAL estimates the Degree of Social Backwardness (DSB) at the urban Basic Geostatistical Area level for 2010 and 2020 based on information from the 2010 and 2020 INEGI Population and Housing Censuses. This information complements the analysis using the data available at the entity, municipality, and locality levels. Based on the DSB information, the classification of the different geographic units is generated in one of the five Degrees: very low, low, medium, high, and very high. In this way, the Social Backwardness Index and the Degree of Social Backwardness allow for identifying priority areas regarding social development for public policies.<sup>7</sup>

Using the CONEVAL database on SBI and DSB, we apply novel techniques such as machine learning combined with Bayesian analysis. Therefore, machine learning is applied (Brunori & Neidhöfer, 2021; Brunori et al., 2023) while complemented with Bayesian learning analysis (Tsagris, 2021) to determine in a non-parametric way how the data and its characteristics indicate what is the primary policy or variable

<sup>7</sup>CONEVAL is a Mexican public agency that coordinates the evaluation of the National Social Development Policy and other policies, programs, or interventions related to social development. See: <https://www.coneval.org.mx/Paginas/principal.aspx>. The SBI is a weighted measure that summarizes indicators of social deprivation (education, health, basic services for homes and home spaces) in a single index and aims to rank observation units according to their social needs. This index generates information for social policy decision-making. It is beneficial for analyzing inequality in social coverage. The SBI is calculated with variables associated with education, access to health services, basic housing services, housing quality, and household assets. These variables consist of the proportion of illiterate population aged 15 years and older, population aged 6 to 14 years who do not attend school, population aged 15 years and older with incomplete primary education, population without the right to health services, homes with ground floors, homes that do not have toilets, homes that do not have running water from the public network, homes that do not have drainage, homes that do not have electricity, homes that do not have a washing machine and homes that do not have a refrigerator. SBI estimates are based on the 2000, 2010, and 2020 Population and Housing Censuses, the 2005 Second Population and Housing Census, and the 2015 Intercensus Survey of the National Institute of Statistics and Geography (INEGI). See: <https://en.www.inegi.org.mx/programas/ccpv/2020/>

to determine and reduce social backwardness, in addition, to have knowledge on the probability that social backwardness reduces.

The remainder of the paper proceeds as follows. Section 2 details the data analysis and the methodologies. Subsection 2.2 applies the “Feature Importance concept” in machine learning to measure the contribution of each variable (XGBoost, Neural Network, Random Forest implementations, and permutation Feature Importance) to the predictive performance of the SBI model. Section 3 presents the Bayesian-ML results, exploratory analysis, and main indicators. Section 4 concludes.

## 2 Methodology and Data Exploration

Our proposal explores machine learning (ML) techniques to analyze data from indicators integrating social backwardness measurement in Mexico. Data used here comes from CONEVAL (2023), i.e., the Social Backwardness Index (SBI), which is a measure that incorporates indicators of educational gap, access to health services, quality, spaces, and basic services in the home and assets in the house. The authors of this research did not develop the indicator; we only describe it in detail below. The construction of the social backwardness index aims to meet three essential criteria. Firstly, incorporate the available information according to the poverty indicators and the levels of disaggregation established by the Law and the available information. Secondly, a database whose structure would allow obtaining indicators at the local, municipal, state, and national levels of aggregation was selected. Based on these first two criteria, we used the database “Main Results by Locality, 2005” of the II Population and Housing Count (ITER 2005). Finally, the principal components statistical technique (PCA) was chosen since it allows the different dimensions of the phenomenon under study to be summarized in an aggregate indicator. Hence, the Social Backwardness Index (SBI) is constructed as a weighted sum of indicators described below. For its construction, the coefficients of the first component (from PCA) are used as weights. Finally, the SBI is standardized, i.e., its mean equals zero, and its variance is unitary. Let us briefly describe the SBI’s indicators, i.e.:

1. Illiteracy Indicator (II):

$$II = \frac{III\_p\_15}{p\_15} \times 100 \quad (1)$$

where,  $III\_p\_15$  is the illiterate population aged 15 years and over, and  $p\_15$  is the population aged 15 years and over.

2. No attending school (NAS):

$$NAS = \frac{nas\_p\_6to14}{p\_6to14} \times 100 \quad (2)$$

where  $nas\_p\_6to14$  is the population from 6 to 14 years old that does not attend school, and  $p\_6to14$  is the population from 6 to 14 years old.

## 3. No Basic Education (NBE):

$$\text{NBE} = \frac{p\_ws + p\_ibs}{p\_ws + p\_ibs + p\_cbs + p\_pbs} \times 100 \quad (3)$$

where  $p\_ws$  is the population aged 15 years and over without schooling, and  $p\_ibs$  is the population aged 15 years and over with incomplete basic education.  $p\_cbs$  is the population aged 15 years and over with complete basic education.  $p\_pbs$  is the population aged 15 years and over with post-basic education.

## 4. No access to health services (NHS):

$$\text{NHS} = \frac{nh\_p}{t\_p} \times 100 \quad (4)$$

where  $nh\_p$  is population without the right to public health services and  $t\_p$  is total Population.

## 5. Dwelling has dirt floor (DF):

$$\text{DF} = \frac{iwdfh}{ih} \times 100 \quad (5)$$

where  $iwdfh$  is inhabited private homes with dirt floors, and  $ih$  is inhabited private homes.

## 6. No toilet (NT):

$$\text{NT} = \left( 1 - \frac{iphwt}{ih} \right) \times 100 \quad (6)$$

where  $iphwt$  is inhabited private homes that have a toilet, and  $ih$  is inhabited private homes.

## 7. No piped water (NPW):

$$\text{NPW} = \frac{iphnp}{ih} \times 100 \quad (7)$$

where  $iphnp$  is inhabited private homes that do not have piped water from the public network, and  $ih$  is inhabited private homes.

## 8. No drainage service (NDS):

$$\text{NDS} = \frac{iphnd}{ih} \times 100 \quad (8)$$

where  $iphnd$  is inhabited private homes that do not have drainage, and  $ih$  is inhabited private homes.

## 9. No electricity (NE):

$$NE = \left(1 - \frac{\text{iphwe}}{\text{ih}}\right) \times 100 \tag{9}$$

where iphwe is inhabited private homes that have electricity, and ih is inhabited private homes.

10. No washing machine (NWM):

$$NWM = \left(1 - \frac{\text{iphww}}{\text{ih}}\right) \times 100 \tag{10}$$

where iphww is inhabited private homes that have a washing machine, and ih is inhabited private homes.

11. No refrigerator (NR):

$$NR = \left(1 - \frac{\text{iphwr}}{\text{ih}}\right) \times 100 \tag{11}$$

where iphwr is inhabited private homes that have a refrigerator, and ih is inhabited private homes.

The indicators above are available for each municipality in Mexico and belong to the years 2000, 2005, 2010, 2015, and 2020 (see, CONEVAL (2023)). Now, let us perform an exploratory data analysis, studying the distribution of the most recent data through histograms. Results are presented in Fig. 1 and Histograms 2. In Figure 1, we notice that in recent years (2020), many Mexican municipalities have moved from high and very high to low and very low social backwardness. Moreover, the number of municipalities at a medium degree does not vary from the other degrees.

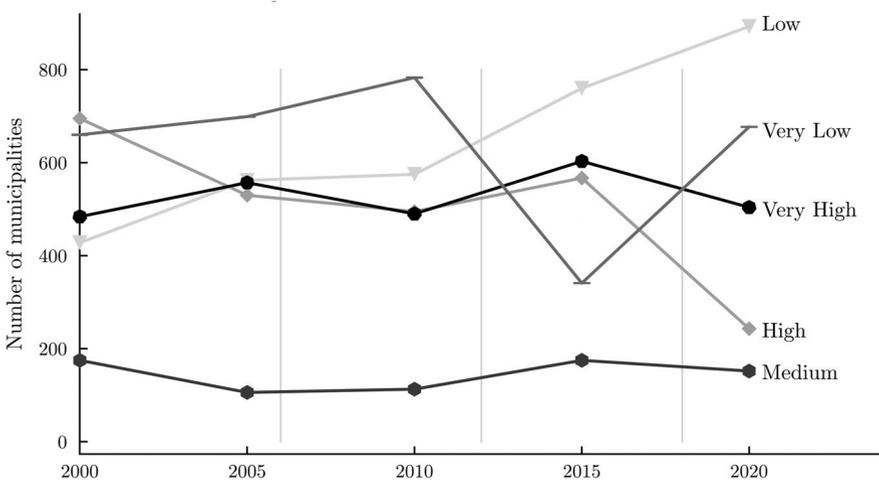
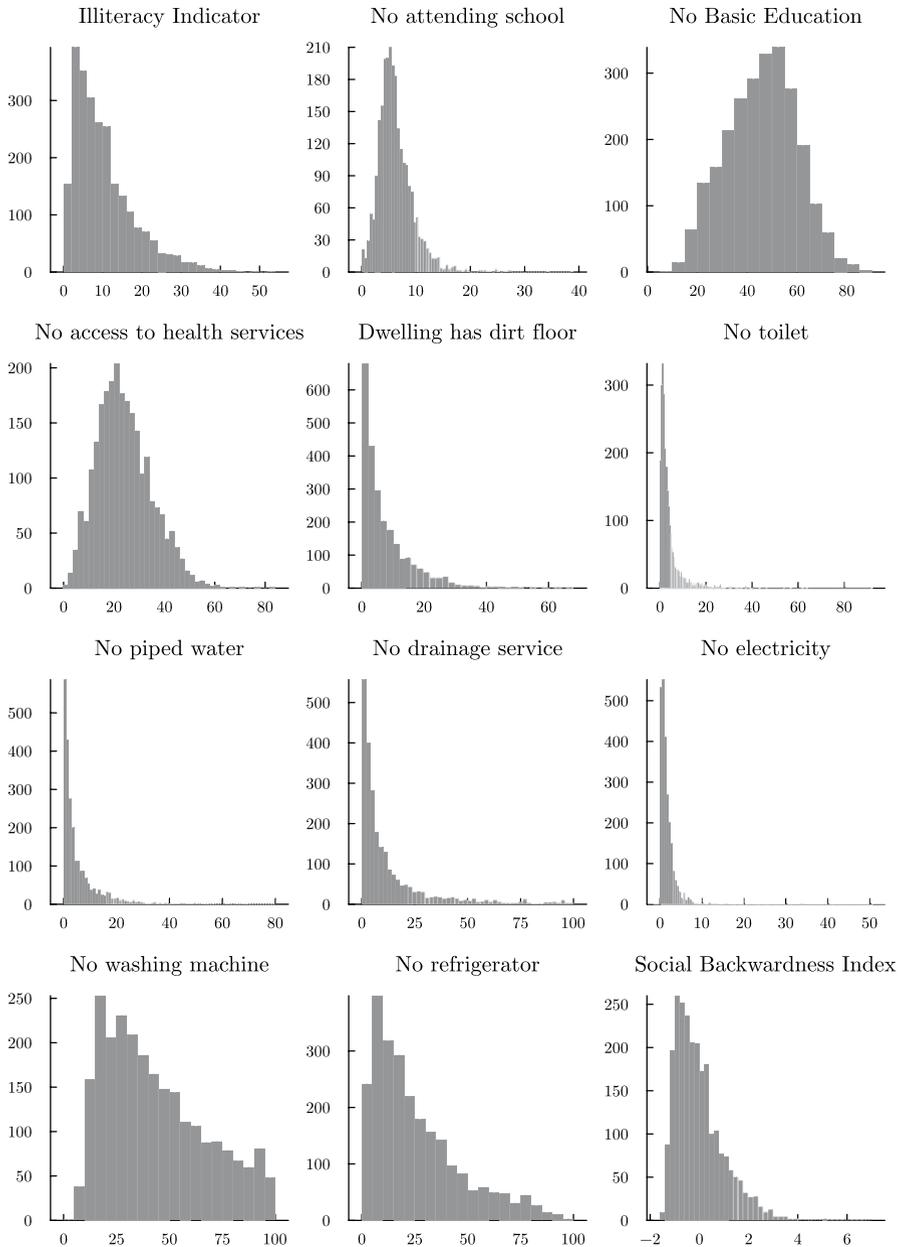


Fig. 1 Degree of Social Backwardness. Number of Municipalities and Years



**Fig. 2** Histograms for each variable considered in 2020, categorized by municipalities, are shown (the x-axis represents the indicator's value, and the y-axis shows the frequency). Several indicators, such as No Toilet, Dwelling has a Dirt Floor, No Piped Water, and No Electricity, exhibit a precise left-skewed distribution. This suggests that most municipalities in 2020 reported lower values for these indicators. Gaussian-like distributions are observed for indicators like No Basic Education and No access to health services

We can verify that from 2000, the year of initial registration of the SBI, until 2020, the indicator has tended to reduce the high and very high positions. It is equivalent to raising the index of the low and very low position; this is mainly due to indicators of health services concerning professional care spaces also in essential services in homes, such as electricity, drinking water, and drainage (Figure 1).

Hence, based on the results of the exploratory analysis, we propose using four widely used ML classification techniques (Random Forests, Multinomial Classification, XGBoost, and Artificial Neural Networks) in addition to a feature selection procedure to identify which variables have the most significant impact on predicting the degree of social backwardness.

## 2.1 Machine Learning Techniques

**Random Forests (RF)** is a supervised learning algorithm proposed by Breiman (2001). RF is an ensemble learning method that combines a series of unrelated  $k$  decision trees,  $T_1, T_2, \dots, T_K$ . One of the advantages of this algorithm lies in reducing the variance by lowering the correlation between trees. RF is trained using the “bagging” method, which combines the bootstrapping and aggregation procedures. Each tree in the forest is formed with a different number of random samples with replacements. Furthermore, several explanatory variables are randomly selected in each tree to perform the partition, and then the tree is built up to a certain depth point. This approach aims to create an improved composite classification or prediction model,  $T^*$ . With a given dataset  $D$ , RF builds multiple decision trees and merges them to get a more accurate and stable prediction. In this algorithm, two of the most important hyperparameters to be considered to perform well in regression tasks are the number of trees in the forest  $N_{trees}$  and the number of split variables  $N_i$  of the preselected root node. In practice, these hyperparameters are tuned empirically. For this reason, it isn't easy to ensure the best performance of Random Forests. However, other approaches, such as Grid search, have been implemented to address this issue. By Liaw and Wiener (2002), for classification purposes, the initial number of features to be considered is  $\sqrt{N}$ , where  $N$  is the number of variables. One of the most interesting characteristics of this technique is that it allows us to know the importance of each input variable. RF measures how feature variations affect the response, known as Feature Importance (FI). In this way, it is concluded that those variables that most significantly influence the variability of the output variable are those that best explain the model. Therefore, these variables can be selected to make the prediction. The importance of the variable  $x_j$  is given by:

$$FI = \frac{1}{N_{trees}} \sum_{v \in S} G(x_j, v) \quad (12)$$

where  $S$  is the set of nodes,  $x_j, v$  is considered to partition the samples and  $G(x_j, v)$  is known as the RF gain of  $x_j$ . Therefore, the gain is based on the impurity measurement when the samples are divided at each node. Several criteria are used to determine the impurity of the data, split it, and determine the degree of importance of a certain feature.

**Multinomial Classification (MC)** is a technique based on the multinomial logistic regression model that generalizes the Linear Regression model by allowing for more than two discrete and unordered response variables. Currently, it is applied in econometrics, psychometrics, engineering, and many other fields, where it uses a set of explanatory variables to predict the probabilities of different possible outcomes of categorically distributed responses (Lin et al., 2014; Abramovich et al., 2021). Suppose  $Y$  is a random variable with a finite number of labeled values of  $1, 2, \dots, L$ . Let

$$p_l = P(Y = l) \text{ and } \sum_{l=1}^L p_l = 1.$$

the objective in MLR be to find a model that relates the probabilities  $p_l$  to the variables  $X_m$ . Similarly,

$$\theta_l = \beta_{0l} + \beta_{1l}X_1 + \dots + \beta_{Ml}X_M = \log \frac{p_l}{p_c}, l = 1, \dots, L - 1,$$

where category  $L$  is used as a reference, and

$$p_l = \frac{e^{\theta_l}}{1 + \sum_{l=1}^{L-1} e^{\theta_l}}, p_L = 1 - \sum_{l=1}^{L-1} p_l.$$

**Extreme Gradient Boosting (XGBoost)** is a model that has been continuously optimized and improved in the follow-up study of many scientists. Unlike Random Forest (Bagging method), XGBoost is a learning framework based on Boosting Tree models (Breiman, 2001; Zhong et al., 2021). This approach performs a second-order Taylor expansion on the loss function. Given a training set  $\{x_i, y_i\}_{i=1}^N$ , a differentiable loss function  $L(y, F(x))$ , a number of weak learners  $M$  and a learning rate  $\alpha$ . A generic XGboost algorithm first initializes a model with constant value:

$$\hat{f}_{(0)}(x) = \arg \min_{\theta} \sum_{i=1}^N L(y_i, \theta)$$

Then for each model  $m \in M$ , compute the gradient and Hessian:

$$\hat{g}_m(x_i) = \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]$$

$$\hat{h}_m(x_i) = \left[ \frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2} \right]$$

now fit the weak learner using the training set  $x_i - \frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}$  by solving the optimization problem:

$$\hat{\phi}_m = \arg \max_{\phi \in \Phi} \sum_{i=1}^N \frac{1}{2} \hat{h}_m(x_i) \left[ -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i) \right]^2$$

$$\hat{f}_m(x) = \alpha \hat{\phi}_m(x)$$

for update the model compute  $\hat{f}_m(x) = \hat{f}_{m-1}(x) + \hat{f}_m(x)$ , and finally the output is:  $\hat{f}(x) = \hat{f}_M(x) = \sum_{m=0}^M \hat{f}_m(x)$ .

**Neural Networks (NN)** are a set of layers that are created by stacking units defined such as  $z_i = \sum_{j=1}^m x_{ij} w_{ij} + b_i$ , where  $m$  is the number of neurons in the current layer,  $w_{ij}$  is the weight of the  $j$ th neuron ( $j$ th input to the  $i$ th cell), and  $b_i$  is a bias term. Given a specific loss function, the perceptron can reach better estimates of the output values by adjusting the weights and bias terms through an iterative process called error correction learning. This technique is enhanced when an activation function is applied to each neuron.<sup>8</sup>

### 2.1.1 Methodology

We now present the steps to follow in the methodology described above.

1. **Preprocessing Data set:** At this stage, the database is examined to ensure that it does not contain missing data and has the appropriate structure for the proposed models. The database is divided into years (2000, 2005, 2010, 2015, 2020). Finally, the variables under study are standardized using Equation (13) to prevent the different magnitudes affecting the performance of the ML algorithms.

$$z = \frac{x - \mu}{\sigma} \tag{13}$$

Where  $x$  is any data,  $\mu$  is the median of some data set, and  $\sigma$  is the standard deviation.

2. **Fitting the ML models:** At this stage, each of the datasets considered in this research (2000 to 2020) is put through the 10-fold cross-validation procedure to fit the four machine learning approaches. This means that ten different models are constructed for each ML algorithm (four ML algorithms) to observe their performances and reduce over-fitting as much as possible. This method splits the dataset into 10 equal-sized sub-samples. Then, a single subsample is retained as the validation data for testing the classification model, and the remaining nine

<sup>8</sup>One example of this non-linear function is the sigmoid, which allows a neural network to perform classification tasks. For details, see Mollalo et al. (2020).

sub-samples are used as training data. Since hyper-parameter tuning is a very important task to enhance the model classification capabilities, a Grid Search procedure is applied to find the hyper-parameters that allow us to fit the datasets better. For details of this procedure, see Mesafint and D H (2021).

3. **Feature Importance:** The importance of indicators is calculated to determine the degree of social backwardness in machine learning that refers to the measure of the contribution of each variable (a.k.a. feature) in an interesting dataset to the predictive performance of a model. In our context, it can be interpreted as determining the most important indicators to determine the degree of social backwardness for each Mexican municipality in a given year or period. Feature Importance can be computed using different ML techniques. Decision Tree-Based Methods (using the impurity) are the widely used techniques in this regard (Breiman, 2001), mainly because these techniques provide Feature Importance scores based on the structure of decision trees (Such as RF and XGBoost). Some authors point out drawbacks to the importance of variables based on impurities since these are biased towards high cardinality characteristics. On the other hand, we are interested in comparing ML methods that do not have the capability of giving Feature Importance (Multinomial regression and Neural Networks). However, to perform this task, we use the Permutation Feature Importance proposed by Altmann et al. (2010), a heuristic method to correct biased measures of Feature Importance.
4. **Bayesian Analysis:** At this stage, we adopt a Bayesian approach to approximate the joint probability distribution and identify the conditional dependencies among variables (the backwardness-related indicators). The objective is to uncover the relationships between these variables, providing insights and tools to propose strategies for mitigating or reducing social backwardness in Mexico. The methodology follows the approach outlined in Sucar (2021) and Tsagris (2021). Specifically, all indicators, except for the backwardness degree indicator, are treated as variables. Structure learning algorithms (such as H2PC and PCHC) are then employed to determine which indicators exhibit statistical dependencies.

Below, we present our results. To our knowledge, this is the first study to apply machine learning and Bayesian network techniques to address the question of the main indicators that explain the social backwardness of Mexican households. Other articles have made statistical inferences based on simple regression models (see, for example: (Benita, 2016; Andrés-Rosales et al., 2018)). Therefore, comparing the results of these studies is irrelevant, as their methodology differs dramatically. Our results are not based on the simple assumptions of standard techniques, such as the model's functional form or the normal distribution of the data.

## 2.2 Results on ML Algorithms and Important Indicators

Table 1 presents the percentage of accuracy for each year in the dataset, and it can be observed that the Neural Network classifier can obtain a better, more stable accuracy percentage compared to a classical classifier known as multinomial classifier (Izenman, 2009). It is worth mentioning that NN has obtained good average accuracy val-

**Table 1** Average and standard deviation of accuracy among classifiers predicting degree of social backwardness (using 10-fold cross-validation)

| Year | Random Forest | Multinomial Classifier | XGBoost          | Neural Network    |
|------|---------------|------------------------|------------------|-------------------|
| 2000 | 79% ± 2.46    | 72% ± 8.26             | <b>91% ± 1.8</b> | 84% ± 2.16        |
| 2005 | 88% ± 1.87    | 74% ± 4.48             | 87% ± 2.18       | <b>91% ± 1.87</b> |
| 2010 | 88% ± 1.31    | 85% ± 10.81            | 89% ± 1.86       | <b>97% ± 1.08</b> |
| 2015 | 91% ± 1.25    | 84% ± 8.90             | 91% ± 1.74       | <b>97% ± 1.12</b> |
| 2020 | 91% ± 1.86    | 92% ± 4.53             | 91% ± 2.45       | <b>96% ± 1.73</b> |

ues (over 90%) through the years, which are acceptable precision values (Tharwat, 2020). While the rest of the approaches are below 80%, with the classical regression approach being the lowest.

These accuracy values were achieved after tuning the hyper-parameters for the four algorithms using the Grid Search technique, shown in Table 2. In addition, the 10-fold cross-validation approach was implemented to avoid overfitting. These tasks represent a computational cost that can become high depending on the model, especially if many hyperparameter combinations are explored and large data sets are used. In summary, while grid search and cross-validation are powerful tools for improving the performance of machine learning models, they also require careful consideration of model resources and complexity. It is important to balance the pursuit of optimal performance with model efficiency and interoperability. The complexity of the ML models used in this work is depicted in Table 3 for the reader’s consideration.

Where  $n$  is the number of training examples,  $m$  is the number of features,  $t$  is the number of trees,  $d$  is the depth of the tree,  $k$  is the number of classes (classification problems), and  $l$  is the number of layers.

**Table 2** Hyper-parameters selected using K-fold Cross Validation and Grid Search

| ML Technique | Hyper-parameter          |                          |                    |
|--------------|--------------------------|--------------------------|--------------------|
| RF           | max depth=20             | n estimators = 500       |                    |
| MC           | multi class= multinomial | solver = lbfgs           |                    |
| XGBoost      | learning rate = 0.5      | max depth = 3            | n estimators = 500 |
| NN           | hidden layer sizes = 100 | learning rate = constant | solver = adam      |

**Table 3** Time and Space Complexity of Different Algorithms

| Algorithm             | Train Time Complexity                | Test Time Complexity   | Space Complexity       |
|-----------------------|--------------------------------------|------------------------|------------------------|
| Logistic Regression   | $O(n \cdot m)$                       | $O(m)$                 | $O(m)$                 |
| Random Forest         | $O(t \cdot n \cdot \log(n) \cdot d)$ | $O(m \cdot t)$         | $O(t \cdot m \cdot d)$ |
| XGBoost               | $O(n \cdot \log(n) \cdot m)$         | $O(m \cdot k)$         | $O(m \cdot k)$         |
| Multilayer Perceptron | $O(n \cdot m \cdot k \cdot l)$       | $O(m \cdot k \cdot l)$ | $O(k \cdot l)$         |

**Table 4** Kruskal-wallis test p-values

| Year | P-value     | significant |
|------|-------------|-------------|
| 2000 | 5.4253e-06  | YES         |
| 2005 | 0.0013      | YES         |
| 2010 | 4.46619e-05 | YES         |
| 2015 | 5.7078e-05  | YES         |
| 2020 | 0.0001      | YES         |

Additionally, to test that the results between the different algorithms are statistically significant, the Kruskal-Wallis test (a nonparametric alternative to one-way ANOVA) is performed since it does not make assumptions about the distribution of the data and is suitable for non-normally distributed data. Table 4 shows the p-values of the result of the Kruskal-Wallis test for the different years, which, being too small (less than 0.05), indicate that the differences between at least one algorithm are significant.

Furthermore, to observe the variability in the predictive capacity of the machine learning models, test datasets that have not been part of the training phase will be used to calculate the accuracy. These metrics will be used to calculate the 95% confidence intervals (CI) using Equation 14, i.e.

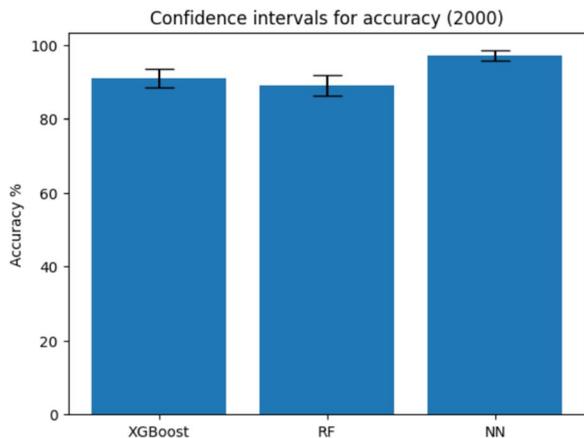
$$CI = z \sqrt{\frac{1}{n} ACC_{test} (1 - ACC_{test})} \tag{14}$$

Where  $z$  is the  $z$  value (the number of standard deviations that a value lies from the mean of a standard normal distribution),  $ACC_{test}$  is the classification accuracy for the test datasets, and  $n$  is the number of samples.

Figure 3 shows the confidence intervals of the models for the year 2000, and Table 5 presents the accuracy percentages and confidence intervals for all years considered in this study. NN demonstrates the highest accuracy percentages in the data tests, with small confidence intervals. This indicates a greater capacity for generalization in predicting and managing uncertainty with this model.

Finally, a Permutation Feature Importance approach is implemented to know which variables were more important in the learning process of the ML techniques.

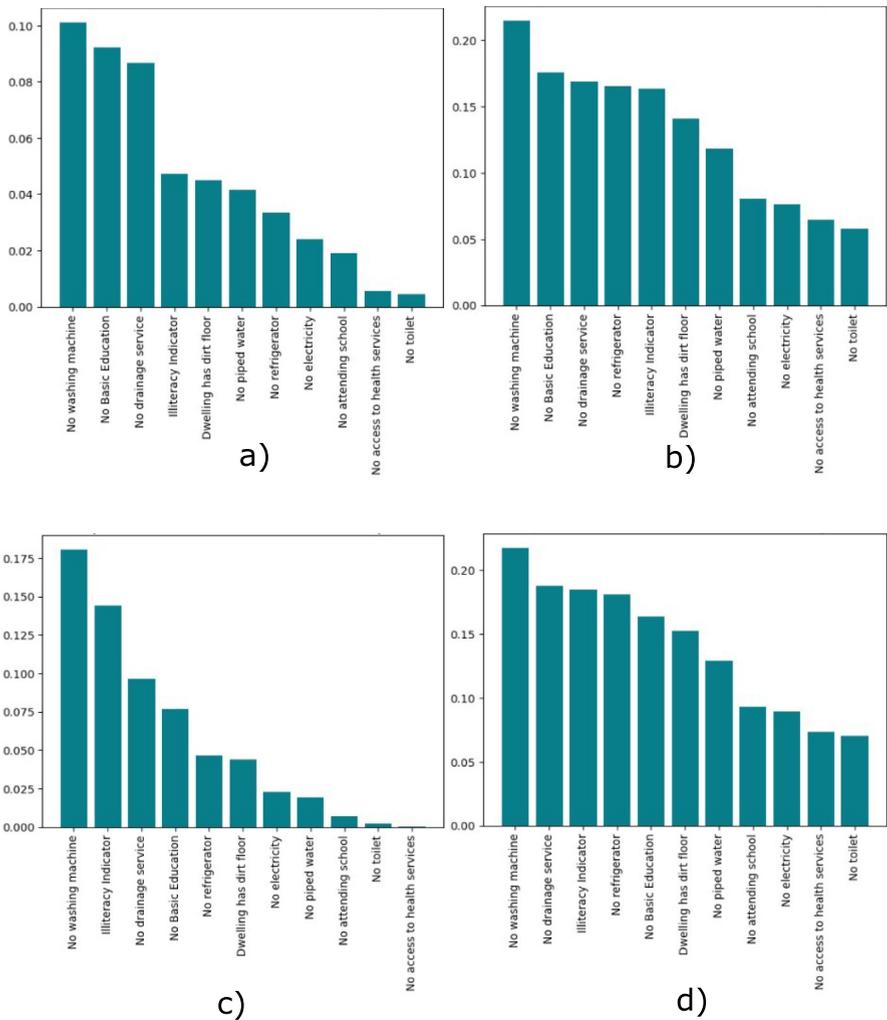
**Fig. 3** Confidence Intervals for 2000



**Table 5** Accuracy test ± Confidence Intervals for ML models

| Year | Random Forest | XGBoost   | Neural Network | number of samples |
|------|---------------|-----------|----------------|-------------------|
| 2000 | 89% ± 2.7     | 91% ± 2.5 | 97% ± 1.4      | 489               |
| 2005 | 87% ± 2.9     | 90% ± 2.7 | 91% ± 2.5      | 491               |
| 2010 | 88% ± 2.8     | 87% ± 2.9 | 97% ± 1.3      | 492               |
| 2015 | 91% ± 2.5     | 91% ± 2.5 | 98% ± 1.2      | 490               |
| 2020 | 90% ± 2.6     | 90% ± 2.6 | 96% ± 1.7      | 494               |

Figure 4 shows the Feature Importance computed for 2010. As can be seen, the four approaches share some variables with major importance, such as “No washing machine”, “No basic education”, “No drainage service”, etc. However, considering



**Fig. 4** Feature Importance in 2010. a) XGBoost, b) Multinomial Regression, c) Random Forest and d) Neural Network

that NN achieved better fitting results, the variables of importance of this technique will be analyzed in depth over the years (2000, 2005, 2010, 2015, 2020) to know how it impacts the SBI.

Next, Figure 5 shows the ranks of each indicator in 2000, 2005, 2010, 2015, and 2020. It can be observed that indicators “dwelling has dirt floor”, “no refrigerator” and “illiteracy indicator” are the most important in 2000. In 2010, RF suggested that the “no washing machine” is the most crucial indicator (25% of importance), followed by the “illiteracy indicator” and “no refrigerator” indicators, both with 17% of importance. In a recent year (2020), the “illiteracy indicator” became the most important indicator, followed closely by “no refrigerator” and “no drainage service” indicators. To classify Mexican municipalities regarding social backwardness, we consider indicators like “no washing machine”, “illiteracy indicator”, “no refrigerator”, and “dwelling has dirt floor”. On the other hand, less important indicators are “no electricity”, “no toilet”, “no attending school” and “no access to health services”. The aforementioned is subject to the accuracy obtained and reported in Table 1. Regarding the essential indicators of social backwardness, the SBI indicates that from 2000 to 2020, educational aspects have reduced their prevalence in said social condition (Figure 5); that is, more people have primary education and fewer abandoned education; however, the illiteracy indicator has raised its degree of importance in the SBI. Population growth has yet to be accompanied by strategies for supplying basic services, given that drainage and dirt floors have increased their hierarchy due to social backwardness from 2010-2020 (from 0.07 to 0.13). Therefore, electricity and sanitary bathrooms have practically remained constant (approximately 0.05 and 0.03), although low (less than 0.05), as characteristics of backwardness, in the same way, health services. As household items, the washing machine characteristic reduced its importance (from 0.25 to less than 0.20), and the refrigerator remained constant at a medium level (0.15).

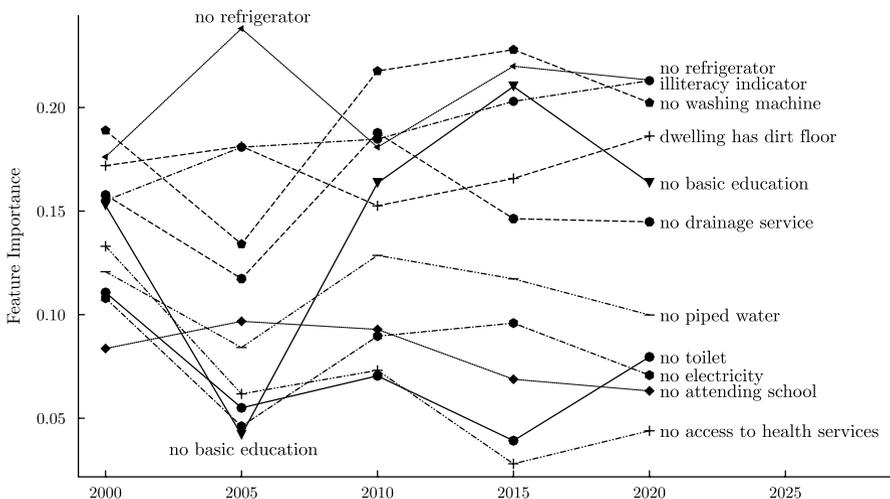


Fig. 5 Feature Importance over the years 2000-2020

From Figure 5, it might be surprising that not having a washing machine at home is a crucial variable that determines social backwardness, as much as it indicates illiteracy. The illiteracy indicator alone explains social backwardness. However, not having a washing machine deserves a brief and detailed analysis. According to the National Survey of Household Income and Expenses (ENIGH) 2020, in Mexico, 11 million homes lack a washing machine. Why is this important, and why does it have to do with social backwardness? The lack of access to household appliances is related to social backwardness because it implies that unpaid work is even more difficult for those who perform these tasks, and in addition, the time to take opportunities in the formal labor market is drastically reduced. In this case, in particular, homemakers suffer from social backwardness due to the lack of washing machines, which affects the home as a whole. According to the National Time Use Survey (ENUT) 2019, while women over 12 years of age who carry out unpaid work cleaning clothes at home dedicate an average of 5 hours per week to this activity, this implies that it leaves them socially behind to undertake any job opportunity.

We have discussed how each variable impacts classification accuracy. However, knowing which variables statistically depend on the others may be helpful. Next, we employ Bayesian network tools to indicate the relationship between variables that make up a specific reality about causes and effects on periodic outcomes over time. The following section studies a Bayesian network approach to approximate the joint distribution of the data and visualize the relationship between indicators, including the Social Backwardness Index.

### 3 Bayesian Analysis Results

Let us apply a Bayesian approach to approximate the joint probability distribution and identify which variables conditionally depend on others. The objective is to find information about the variables and how they relate to each other to provide tools that suggest possible strategies to mitigate or reduce social backwardness in Mexico.

#### 3.1 Dependence Between Indicators and Social Backwardness Index

Based on Bayesian probability theory, Bayesian learning is a framework for statistical modeling and ML (Sucar, 2021). It offers a logical and probabilistic method for modeling and producing forecasts where there is uncertainty. The fundamental tenet of Bayesian learning is that model inputs and predictions should be viewed as probability distributions instead of fixed values. Here, we use Bayesian Networks (BNs) to represent dependency between observed variables. BNs are considered probabilistic graphical models that effectively and factorized depict the dependency structure of a group of variables and their joint distribution. We use the *bnlearn* package using R language (Nagarajan et al., 2013) to compute the BNs for each interest year. We select the structure learning algorithm known as H2PC, which is a hybrid algorithm combining HPC and a hill-climbing optimizer. It is worth mentioning that this structured learning method provided simple and useful networks to describe social backwardness. Figures 6-10 show the resulting Bayes network for each five years from

2000 to 2020. Bayesian networks indicate the cause and effect (cause  $\rightarrow$  effect) between two or more variables; that is, the arcs of the network indicate statistical dependence, e.g.,  $A \rightarrow B$  means that  $B$  conditionally depends on  $A$  (the probability whether  $B$  occurs depends on whether evidence  $A$  occurs). In terms of joint probability,  $P(A, B) = P(B|A) \times P(A)$ .

In Figure 6, in 2020, the least relevant and independent variables are “no attending school” and “no access to health services”. Regarding the importance of the SBI, the variable “lack of a washing machine” is crucial in social backwardness, as it has the effect of not having a refrigerator, which directly affects the SBI. Then, the SBI goes into five variables as dependency effects in the Bayesian scheme, with not having a basic education being the final SBI effect.

That is, Figure 6 shows that the indicators “no washing machine”, “no toilet,” and “no attending school” are statistically independent variables. However, the rest of the variables depend conditionally on each other. For example, the “Social Backwardness Index” conditionally depends on the indicators “no washing machine”, “no toilet” and “no refrigerator”. While the “illiteracy indicator” conditionally depends on the backwardness index”. Furthermore, the indicator of “no access to health services” depends on “no attending school”. That is to say, with social backwardness, there is illiteracy (the latter depends on the former, not vice versa). When you do not have access to health services, this is conditionally due to not attending school. The social backwardness is because there is no toilet, refrigerator, or washing machine.

In Figure 7, in the year 2015, we get that three indicators are independent: “no washing machine”, “no attending school,” and “no toilet.” Notice that the “social

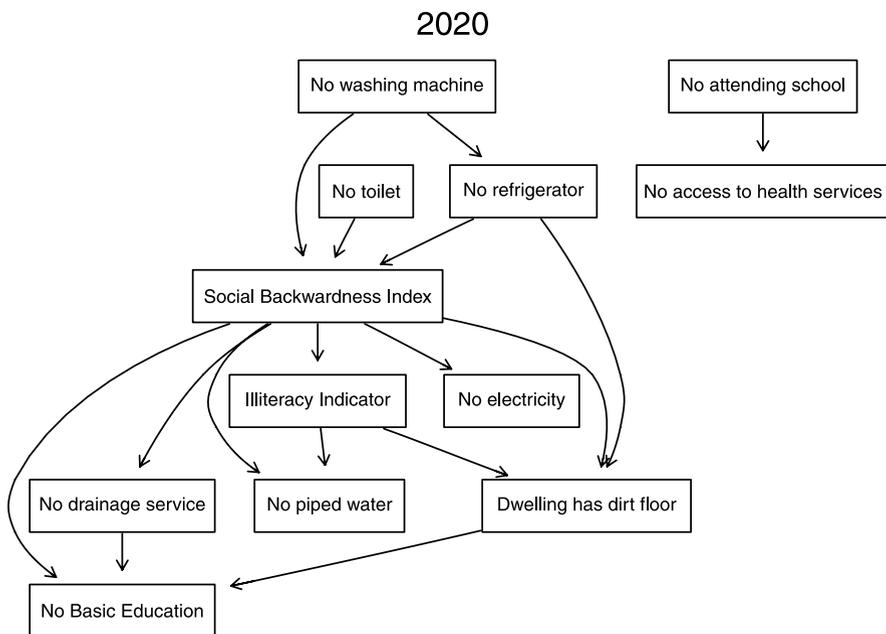


Fig. 6 Resulting Bayes network for 2020

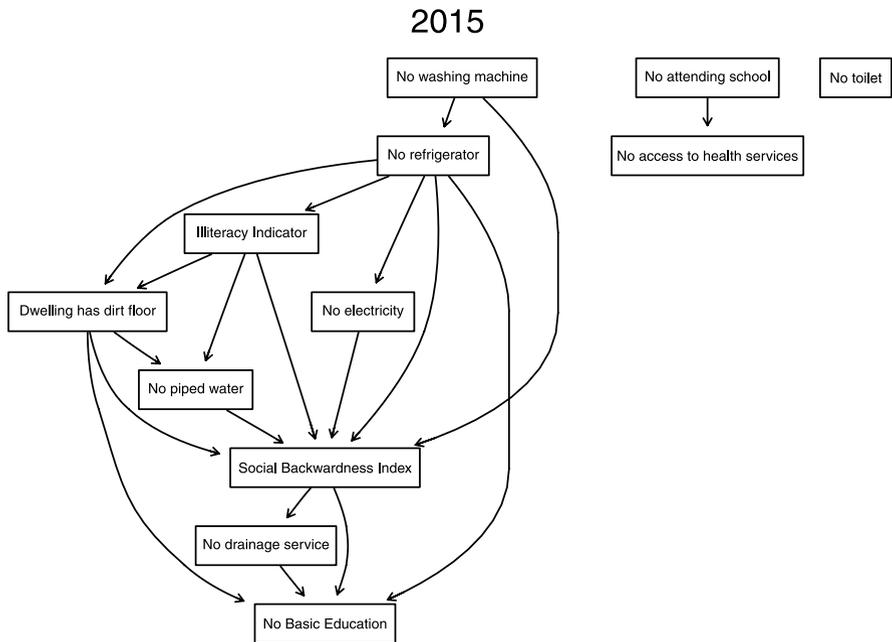


Fig. 7 Resulting Bayes network for 2015

backwardness index” is evidence of an effect on the indicators “no drainage service” and “no basic education”, but most of the indicators are having an impact on the "social backwardness index." Additionally, “no toilet” is an isolated variable for the 2015 data. As it is in 2020, social backwardness depends on not having basic education and a lack of household appliances and basic services at home.

In 2010, school attendance and bathroom installation were independent and unimportant to the SBI. On the other hand, not having a refrigerator is essential, and it has a dependency effect on not having health services and not having electricity, although these two are also causes. The remaining five variables are less relevant as fate-effects of the analytical scheme, with dirt floor as the final effect (Figure 8).

Figure 9 shows that, in 2005, the “social backwardness index” only depends on the indicators of “no toilet”, “no refrigerator”, and “no washing machine”. Still, the same index is affecting most of the remaining indicators.

In 2000, school attendance and electricity were isolated or independent variables irrelevant to the SBI. Dependency and prevalence are related to the lack of washing machines and refrigerators. On the other hand, illiteracy (no attending school) and dirt floors are shown to be the ultimate effects of dependency coming from the SBI. (Figure 10).

Figure 10 shows that the “social backwardness index” only conditionally depends on the “no refrigerator” indicator. However, the mentioned index is causing effect to “no piped water”, “No Basic Education”, “No access to health services”, “No toilet”, “No drainage service N” and “No electricity”. Moreover, the “no washing machine” and “no attending school” indicators are statically independent, and the last one is not influenced by or does not influence any other variables in the network.

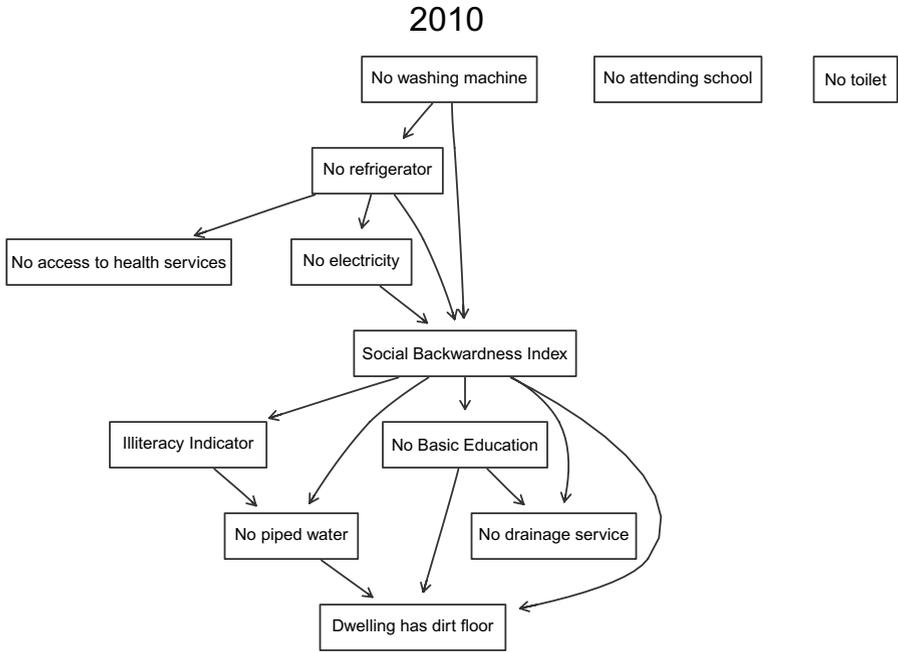


Fig. 8 Resulting Bayes network for 2010

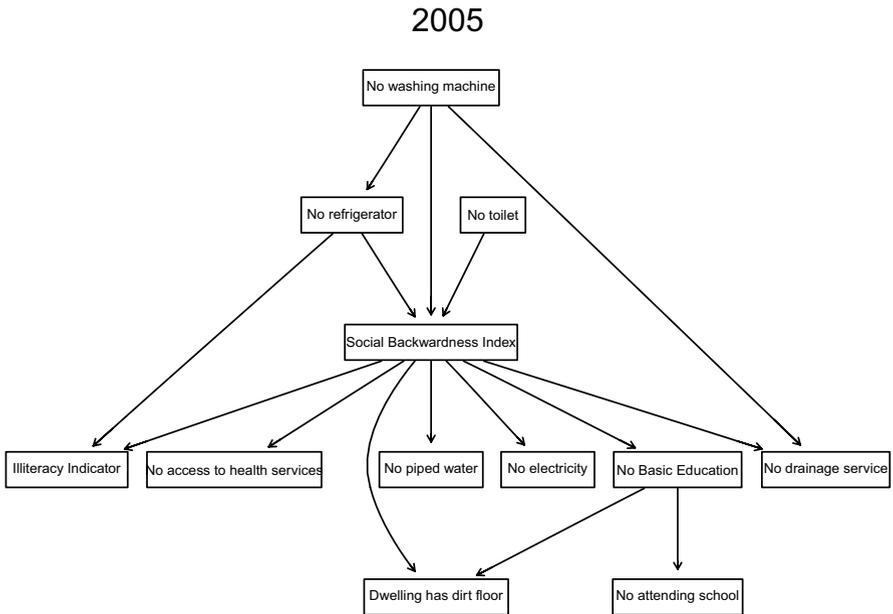


Fig. 9 Resulting Bayes network for 2005

2000

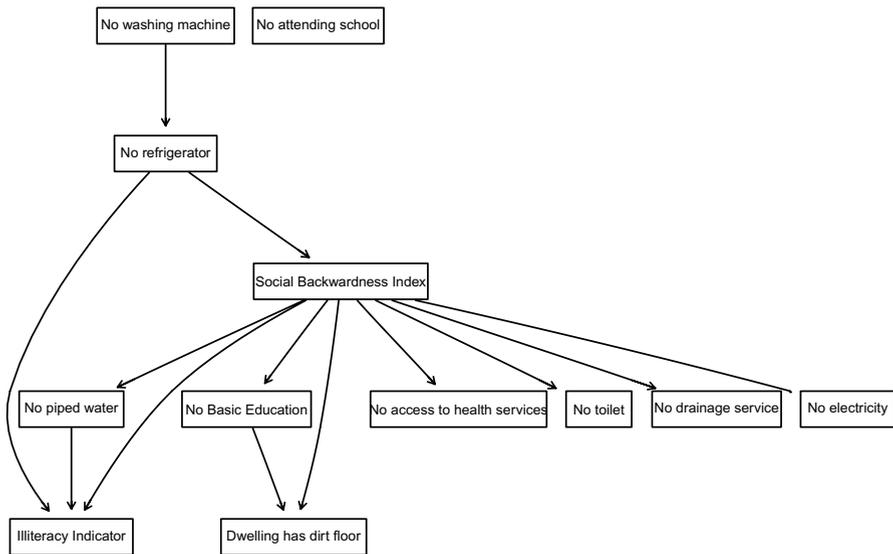


Fig. 10 Resulting Bayes network for 2000

It is important to note that our results imply dependencies and effects from several indicators on the SBI. Hence, it is worth noting that although Bayesian networks can reveal dependencies between variables, these relationships are not equivalent to causal relationships. Factors such as “lack of washing machines” have strong predictive power of social backwardness, but we do not fully demonstrate whether this association implies causality. For example, the lack of washing machines could be an external manifestation of social backwardness, rather than its root cause.

Therefore, economic policies aimed at improving the services and assets of Mexican households will contribute to reducing their social backwardness. In other words, economic policies aimed at addressing social backwardness should promote a combination of highly targeted social welfare programs, affirmative action, and social reform initiatives at the household level (Shah, 1991; Sharma & Paramati, 2018). These economic policies should seek to empower individuals and communities facing social disadvantage by providing them with access to essential household services, education, and economic opportunities, while also addressing systemic inequalities.

Promoting, for instance, a policy on the provision of basic physical infrastructure for households must be implemented as a means of combating social backwardness and poverty. Infrastructure refers to durable resources, consisting of collective goods that generate externalities. This includes the health and education of household members. Its relevance varies depending on the geographic scope of the study, as it involves factors such as location, economic, industrial, and social development, climate, raw materials, customs, and public policies. Thus, infrastructure provides a secure means of providing services; its access can be affordable; it stimulates the economy, and includes marginalized groups. Additionally, it reduces the cost of access to markets,

increases surplus value, encourages the accumulation of human capital, and enables the dissemination of knowledge, which requires structured, harmonious processes equipped with the necessary economic, material, human, and intellectual resources.

However, great caution must be exercised in the recommendations and implementation of economic policies to reduce social backwardness. Social policy has become a poverty alleviation tool and does not promote equal opportunity. According to Coneval Mexico, there are 6,491 social development programs across the country, but in addition to many overlapping programs, not all of them aim to eradicate poverty (Campos & Trigueros, 2023). Oxfam proposes an approach based on universal social rights, consolidating an effective universal health system and ensuring that social programs have a rights-based approach. It also aims to evaluate the viability of a pilot program for the implementation of the Universal Basic Income.<sup>9</sup>

### 3.2 What is the Probability that Social Backwardness Will be Reduced?

To answer this question, we employ Bayesian inference, a helpful feature the fitted Bayesian network provides to calculate probabilities on given scenarios. Bayesian inference is predicated on revising assumptions regarding an event or hypothesis to calculate the joint probability of the event in question (Sucar, 2021). In what follows, we determine the likelihood that social backwardness would be reduced given that the indicators related to it are taking the most appropriate values (evidence of overcoming specific threshold values).

In this part, we focus on computing the probability associated with recently measured data using a Bayesian network, as illustrated in Fig. 6. The Social Backwardness Index (SBI) is identified as an outcome influenced by the No Washing Machine (NWM), No Toilet (NT), and No Refrigerator (NR) indices. Consequently, our objective is to determine

$$P(SBI|NT, NWM, NR)$$

with a particular emphasis on maximizing this probability when the SBI aligns with low and very low degrees of social backwardness, denoted as  $SBI < -0.6$  according to empirical data. We identify values for the indices NT, NWM, NR that maximize  $P(SBI < -0.6|NT, NWM, NR)$  to achieve this. Given that NT, NWM, and NR are positive indices, we employ a systematic grid search approach, discretizing the interval for each index and recording the corresponding values where the probability is maximized. This computational approach suggests that the social backwardness reduction, indicated by  $SBI < -0.6$ , is most likely achieved when the conditions  $NT \geq 0.0$ ,  $NWM \geq 5.7$ ,  $NR \geq 24.87$  are met, obtaining the maximum probability value in the required scenario:

$$P(SBI < -0.6|NT \geq 0.0, NWM \geq 5.7, NR \geq 24.87) = 0.264.$$

In the best scenario, social backwardness will be reduced with a probability of 0.264 as long as NT, NWM, and NR are in the provided range, perhaps promoted by public

<sup>9</sup><https://oxfamexico.org/mexico-justo-politicas-publicas-contra-la-desigualdad-0/>

policies that increase these indicators. That is to say, household appliances are a primary condition of Mexican homes to reduce social backwardness.

Recall that the Index of Social Backwardness (SBI) measures the degree of progress in well-being in the four dimensions of social development, which are educational lag, access to health services, quality of housing, its spaces, and essential services; it also adds household goods, such as a refrigerator and washing machine. It indicates access to social rights and the availability of goods supporting well-being as a life condition. However, it does not incorporate income indicators, access to comprehensive social security and food.

Finally, to compare the networks (BNs) provided by the package bnlearn, we utilized the PCHC method proposed by Tsagris (2021) to identify Bayesian networks in economic applications. The resulting networks generated by PCHC are displayed in Fig. 11. Notably, PCHC produced more complex network architectures, characterized by more connections among variables, than those obtained using H2PC in bnlearn.

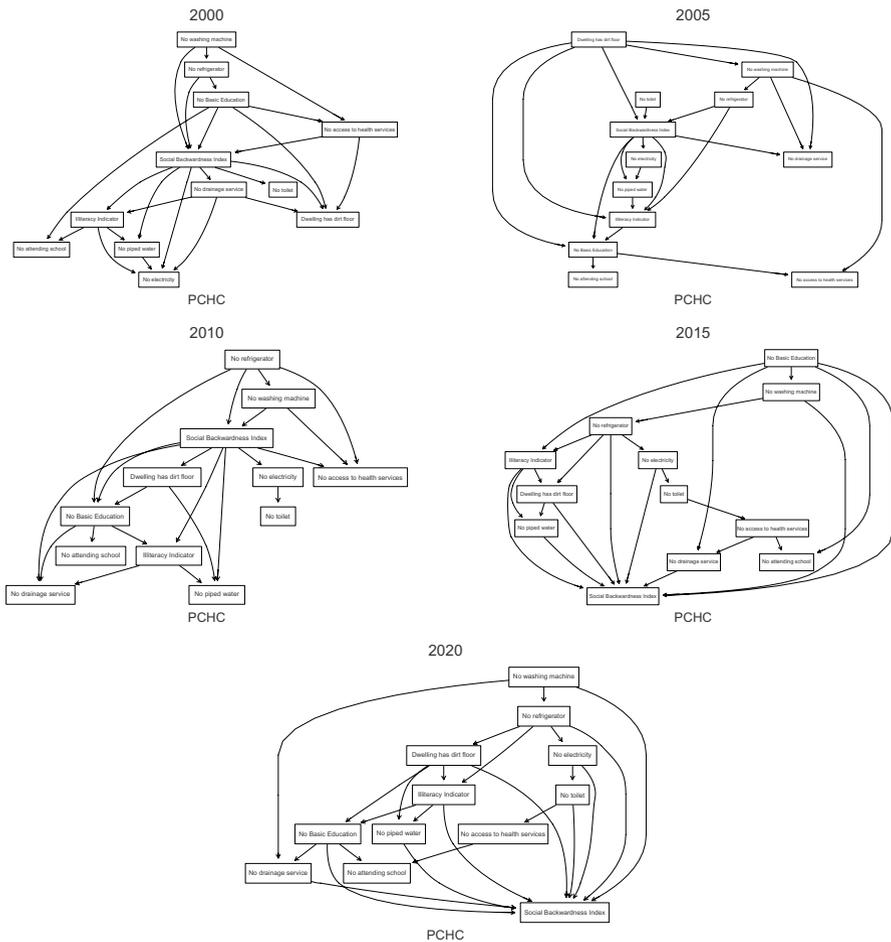


Fig. 11 Resulting Bayes network for 2000 discovered by PCHC algorithm

Overall, Figure 11 confirms the previous results. In 2000, 2005, and 2010, not having a refrigerator or a washing machine and not having basic primary education (in 2015) implied a probability of being socially backward. However, with this other methodology, it is worth noting that for the most recent year, 2020, basic primary education does not explain the social backwardness of Mexican households. With this, we conclude and give the final considerations below.

## 4 Concluding Remarks

In this research paper, Bayesian network machine learning approaches are applied to show the leading indicators that cause and predict the social backwardness of Mexican households. Feature Importance and Bayesian probabilities show that the lack of electronic devices, such as washing machines and refrigerators, would be considered a relevant factor in social backwardness, given that the main cause of poverty is income level, derived from a lack of education and training, as well as a lack of comprehensive healthcare, among other consequences.

Therefore, from the perspective of a backwardness condition, the lack of basic goods that facilitate daily life is, apart from the level of poverty, a condition of lag concerning other sectors of society, such as the classes with moderate poverty and the middle and upper classes. In this sense, Mexico's economic policy regarding the estimation of the SBI shows that in the years indicated, 80.4% of the country's localities (86,949) had improvement. With persistent, high, and very high social backwardness, 5,258 were registered in 2020, representing 74%. According to the SBI results, between 2000 and 2020, the percentage of localities with high social backwardness increased from 31.2 to 4.8 percent, and the rate of localities with very high social backwardness rose from 22.9 to 1.8 percent of the total. In 2020, only 386,154 people lived in localities with high and very high social backwardness, with less access to social rights and less availability of goods in homes, compared to the rest of the localities this year. Thus, in 2020, in Mexico, 11.5 million people lived in high and very high social backwardness, 9.6 and 1.9, respectively. In terms of localities in the country, in 2000, out of 107,215, 33,409 were detected with high lag and 24,503 with very high lag, while in 2020, out of 108,149 with high lag, 5,162 were detected and with very high lag, 1,974. In other words, significant progress is needed. Regarding educational lag or incomplete basic education, 29.6% of the population aged 15 years and older are in that condition, with 4.7% illiteracy, in addition to 6.1% of the population aged 6 to 14 years who do not attend school. According to CONEVAL, in the same year, 2020, the results indicate that 26.2% of the population does not have access to health services, in consumer durables, 27.2% of the population does not have a washing machine, and 12.4% does not have a refrigerator.

By applying the ML tool and Bayesian Networks, it is possible to infer the degree of importance of certain variables (here we exploit eleven indicators related to education, health, and household well-being, in general) that make up a specific reality about the causes and effects on periodic results in a given time to predict social backwardness in Mexico. The above implies that the more education and training characterize the members of Mexican households, the more vulnerable population would

have access to household goods that contribute to well-being and reduce social backwardness, but it is a long-term economic policy. Economic policy and future research should consider variables such as family income, individual skills, age, gender, work experience, etc., to explain how these critical variables cause social backwardness. Essential and crucial also for economic policy and future research is the fact that, for example, what happens if people are provided with washing machines (taking into account that they can maintain the machine -for example, pay for electricity-) but face other problems (more related to institutional issues) such as the level of corruption or crime levels? So, will we get a reduction in social backwardness, or will we see a decline in the index because one of the components of the IBS is reduced? These questions remain open for future research.

**Funding** Open access funding provided by Università degli Studi di Firenze within the CRUI-CARE Agreement.

## Declarations

**Competing Interests** No funds, grants, or other support was received.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abramovich, F., Grinshtein, V., & Levy, T. (2021). Multiclass classification by sparse multinomial logistic regression. *IEEE Transactions on Information Theory*, 67(7), 4637–4646. <https://doi.org/10.1109/IT.2021.3075137>
- Agüero, J. M., & Beleche, T. (2013). Test-mex: estimating the effects of school year length on student performance in mexico. *Journal of Development Economics*, 103, 353–361.
- Altmann, A., Tološi, L., Sander, O., et al. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10), 1340–1347.
- Andrés-Rosales, R., Bustamante Lemus, C., Ramírez Argumosa, G.S. (2018). Social Investigaciones Regionales-Journal of Regional Research, (40), 57–78. <https://ideas.repec.org/a/ris/invreg/0365.html>
- Athey, S. (2018). The impact of machine learning on economics. In: The economics of artificial intelligence: An agenda. University of Chicago Press, p 507–547, <https://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda>
- Balasankar, V., Penumatsa, S., Terlapu, P.V. (2020). Intelligent socio-economic status prediction system using machine learning models on rajahmundry a.p., ses dataset. *Indian Journal of Science and Technology*, 13(37), 3820–3842. <https://doi.org/10.17485/IJST/v13i37.1435>
- Benita, F. (2016). Social backwardness in mexico city metropolitan area. *Social Indicators Research*, 126(1), 141–160. <https://www.jstor.org/stable/48714599>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>

- Brida, J., Sanchez Carrera, E. J., Risso, W. A., et al. (2021). Growth and inequality in the mexican states: regimes, thresholds, and traps. *Papers in Regional Science*, 100(5), 1295–1322. <https://doi.org/10.1111/pirs.12616>
- Brunori, P., Hufe, P., & Mahler, D. G. (2023). The roots of inequality: estimating inequality of opportunity from regression trees and forests. *The Scandinavian Journal of Economics*, 125(4), 900–932. <https://doi.org/10.1111/sjoe.12530>
- Brunori, P., & Neidhöfer, G. (2021). The evolution of inequality of opportunity in germany: a machine learning approach. *Review of Income and Wealth*, 67(4), 900–927.
- Campos, F., Trigueros, J. (2023). Fondo de infraestructura social estatal en méxico: Un análisis de política pública. *Revista Economía y Política*, pp 31–47. <https://doi.org/10.25097/rep.n37.2023.03>
- Chugh, G., Kumar, S., & Singh, N. (2021). Survey on machine learning and deep learning applications in breast cancer diagnosis. *Cognitive Computation*, 13, 1451–1470. <https://doi.org/10.1007/s12559-020-09813-6>
- Ciaburro, G. (2022). Machine fault detection methods based on machine learning algorithms: a review. *Mathematical Biosciences and Engineering*, 19(11), 11453–11490.
- CONEVAL. (2023). Índice de rezago social. Consejo Nacional de Evaluación de la Política de Desarrollo Social. <https://www.coneval.org.mx/Medicion/IRS/Paginas/Que-es-el-indice-de-rezago-social.aspx>
- Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, 346(6210), 1243089. <https://doi.org/10.1126/science.1243089>
- Feldmeyer, D., Meisch, C., Sauter, H., et al. (2020). Using openstreetmap data and machine learning to generate socio-economic indicators. *ISPRS International Journal of Geo-Information*, 9(9). <https://doi.org/10.3390/ijgi9090498>, <https://www.mdpi.com/2220-9964/9/9/498>
- Felix, J., Alexandre, M., Lima, G. (2024). Applying machine learning algorithms to predict the size of the informal economy. *Computational Economics*, pp 1–21. <https://doi.org/10.1007/s10614-024-10593-6>
- Gogas, P., & Papadimitriou, T. (2021). Machine learning in economics and finance. *Computational Economics*, 57(1), 1–4. <https://doi.org/10.1007/s10614-021-10094-w>
- Gonzalez-Jimenez, D., del Olmo, J., Poza, J., et al. (2021). Machine learning-based fault detection and diagnosis of faulty power connections of induction machines. *Energies*, 14(16). <https://doi.org/10.3390/en14164886>, <https://www.mdpi.com/1996-1073/14/16/4886>
- Hassan, A., Muse, A., & Chesneau, C. (2024). Machine learning study using 2020 sdhs data to determine poverty determinants in somalia. *Scientific Reports*. <https://doi.org/10.1038/s41598-024-56466-8>
- Izenman, A. (2009). Modern multivariate statistical techniques: regression, classification, and manifold learning. *Springer Texts in Statistics*, Springer, New York, NY. <https://doi.org/10.1007/978-0-387-78189-1>
- Kwekha-Rashid, A. S., Abduljabbar, H. N., & Alhayani, B. S. A. (2021). Coronavirus disease (covid-19) cases analysis using machine-learning applications. *Applied Nanoscience*, 13, 2013–2025. <https://doi.org/10.1007/s13204-021-01868-7>
- Lazebnik, T. (2024). Going a step deeper down the rabbit hole: Deep learning model to measure the size of the unregistered economy activity. *Computational Economics*, pp 1–16. <https://doi.org/10.1007/s10614-024-10606-4>
- Liaw, A., Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3), 18–22. <http://CRAN.R-project.org/doc/Rnews/>
- Lin, Y., Deng, X., & Li, X. (2014). Comparison of multinomial logistic regression and logistic regression: which is more efficient in allocating land use? *Frontiers in Earth Science*, 8, 512–523. <https://doi.org/10.1007/s11707-014-0426-y>
- Liu, Z., Xiao, C., Zhang, T., et al. (2020). Research on fault detection for three types of wind turbine sub-systems using machine learning. *Energies*, 13(2). <https://doi.org/10.3390/en13020460>
- Mesafint, D., & D H, M. (2021). Grid search in hyperparameter optimization of machine learning models for prediction of hiv/aids test results. *International Journal of Computers and Applications*, 44, 1–12. <https://doi.org/10.1080/1206212X.2021.1974663>
- Mollalo, A., Rivera, K., & Vahedi, B. (2020). Artificial neural network modeling of novel coronavirus (covid-19) incidence rates across the continental united states. *International Journal of Environmental Research and Public Health*, 17,. <https://doi.org/10.3390/ijerph17124204>
- Monroe, B.L., Pan, J., Roberts, M.E., et al. (2015). No! formal theory, causal inference, and big data are not contradictory trends in political science. *PS: Political Science & Politics*, 48(1), 71–74. <https://doi.org/10.1017/S1049096514001760>
- Nagarajan, R., Scutari, M., & Lebre, S. (2013). *Bayesian Networks in R: with Applications in Systems Biology*. <https://doi.org/10.1007/978-1-4614-6446-4>

- Ozden, E., & Guleryuz, D. (2022). Optimized machine learning algorithms for investigating the relationship between economic development and human capital. *Computational Economics*, 60(1), 347–373.
- Polyzos, E., & Siriopoulou, C. (2023). Autoregressive random forests: machine learning and lag selection for financial research. *Computational Economics*, 64, 1–38. <https://doi.org/10.1007/s10614-023-10429-9>
- Radovanovic, J., & Haas, C. (2023). The evaluation of bankruptcy prediction models based on socio-economic costs. *Expert Systems with Applications*, 227, Article 120275.
- Risso, W. A., Punzo, L. F., & Carrera, E. J. S. (2013). Economic growth and income distribution in Mexico: a cointegration exercise. *Economic Modelling*, 35, 708–714.
- Sanchez Carrera, E.J., Risso, W.A. (2023). On Mexican poverty-trap regimes and struggling to escape them. *Macroeconomic Dynamics*, pp 1–29. <https://doi.org/10.1017/S1365100523000275>
- Shah, G. (1991). Social backwardness and politics of reservations. *Economic and Political Weekly*, 26(11/12), 601–610. <http://www.jstor.org/stable/4397417>
- Shami, L., & Lazebnik, T. (2024). Implementing machine learning methods in estimating the size of the non-observed economy. *Computational Economics*, 63(4), 1459–1476. <https://doi.org/10.1007/s10614-023-10369->
- Sharma, C., & Paramati, S. R. (2018). Measuring inequality of opportunity for the backward communities: regional evidence from the Indian labour market. *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement*, 138(2), 479–503.
- Spandagos, C., Tovar Reaños, M. A., & Lynch, M. A. (2023). Energy poverty prediction and effective targeting for just transitions with machine learning. *Energy Economics*, 128, Article 107131.
- Storm, H., Baylis, K., & Heckelesi, T. (2019). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, 47(3), 849–892.
- Sucar, L. E. (2021). Probabilistic graphical models. *Springer International Publishing*. <https://doi.org/10.1007/978-3-030-61943-5>
- Tagliati, F. (2022). Welfare effects of an in-kind transfer program: evidence from Mexico. *Journal of Development Economics*, 154, Article 102753.
- Thalari, S.K., Dileep, P., Latha, Y.M., et al. (2023). Predictive modelling of socioeconomic trends using machine learning: Implications for policy planning. *Journal of Namibian Studies: History Politics Culture*, 35, 2758–2772. <https://doi.org/10.59670/jns.v35i.4126>
- Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168–192. <https://doi.org/10.1016/j.aci.2018.08.003>
- Tsagris, M. (2021). A new scalable Bayesian network learning algorithm with applications to economics. *Computational Economics*, 57(1), 341–367.
- Varian, H. R. (2014). Big data: new tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3–28. <https://doi.org/10.1257/jep.28.2.3>
- Zhang, J., Wang, Z., Hu, J., et al. (2022). Bayesian machine learning-based method for prediction of slope failure time. *Journal of Rock Mechanics and Geotechnical Engineering*, 14(4), 1188–1199.
- Zhong, S., Zhang, K., Bagheri, M., et al. (2021). Machine learning: New ideas and tools in environmental science and engineering. *Environmental Science & Technology*, 55(19), 12741–12754, PMID: 34403250. <https://doi.org/10.1021/acs.est.1c01339>, <https://doi.org/10.1021/acs.est.1c01339>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

Jesús Alejandro Navarro-Acosta<sup>1</sup>  · Jesús-Adolfo Mejía-de-Dios<sup>1</sup>  ·  
José María González Lara<sup>2</sup>  · Edgar J. Sanchez Carrera<sup>1,3</sup> 

✉ Edgar J. Sanchez Carrera  
edgar.sanchez@unifi.it

Jesús Alejandro Navarro-Acosta  
alejandro.navarro@uadec.edu.mx

Jesús-Adolfo Mejía-de-Dios  
adolfomejia@uadec.edu.mx

José María González Lara  
josegonzalezlara@uadec.edu.mx

- <sup>1</sup> Research Center in Applied Mathematics, CIMA UAdeC, Edificio S, Unidad Camporredondo, Saltillo 25280, Coahuila, México
- <sup>2</sup> Facultad de Economía, UAdeC, Edificio E, Unidad Camporredondo, Saltillo 25280, Coahuila, México
- <sup>3</sup> Department of Economics and Management, University of Florence, Via delle Pandette 32, Florence 50127, Tuscany, Italy