Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case

Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

Itzel Coquis Rioja

EGADE Business School, Tecnológico de Monterrey, Mexico

Mario Iván Contreras Valdez

Tecnológico de Monterrey, Mexico

Abstract

The current article explores the causes of financial inclusion among the Mexican population. It leverages data from the Encuesta Nacional de Inclusión Financiera (ENIF) (INEGI, 2021) to develop two machine learning models aimed at identifying individuals who are part of the financial system. These models are assessed using both artificial intelligence methodologies and traditional statistical significance tests. The findings suggest that factors such as education level, monthly income, future-oriented behavioral preferences over present ones, saving capacity, and access to smartphones are significant drivers that enhance the likelihood of financial inclusion. Consequently, there is a potential for implementing public policies to incentivize individuals to voluntarily adopt formal financial services.

Keywords: *financial inclusion, artificial intelligence, machine learning.*
JEL Classification: *C13, C54, D14.*

Resumen

El presente artículo explora las causas de la inclusión financiera entre la población mexicana. Con datos de la Encuesta Nacional de Inclusión Financiera (ENIF) (INEGI, 2021) desarrolla dos modelos de aprendizaje automático con el objetivo de identificar a individuos que forman parte del sistema financiero. Estos modelos son evaluados valiéndose tanto de metodologías de inteligencia artificial como de pruebas estadísticas de significancia tradicionales. Los hallazgos sugieren que factores como nivel educativo, ingreso mensual, preferencias orientadas hacia el futuro sobre las presentes, capacidad de ahorro y acceso a teléfonos inteligentes son impulsores significativos que aumentan la probabilidad de inclusión financiera. En consecuencia, existe un potencial para la implementación de políticas públicas dirigidas a incentivar a los individuos para que adopten voluntariamente servicios financieros formales.

Palabras clave: *inclusión financiera, inteligencia artificial, aprendizaje automático.*
Clasificación JEL: *C13, C54, D14.*

1. Introduction

Throughout human history, technological advancements have catalyzed numerous social changes. Initially, these innovations often impacted only small groups of people; however, as democratization progresses, their benefits become increasingly accessible to broader populations. In the realm of economic and financial interactions, such advancements have led to new modes of transactional engagement. One such innovation is financial inclusion. According to basic economic theory (Varian, 1992), uncertainty detrimentally affects individuals' utility. In such circumstances, the ability to safely transfer wealth across time periods becomes a valuable tool for enhancing societal welfare. Financial markets and institutions facilitate this temporal decision-making process. Despite these advantages and available technologies, many countries have yet to fully exploit them. Demirguc-Kunt and Klapper (2012) define financial inclusion as the capacity of individuals to access financial products and services that meet their needs sustainably. Building upon this definition, Tram et al. (2023) highlight three primary dimensions that determine financial inclusion: penetration, availability, and utilization of financial services.

Given the significance of financial services and markets in modern economies, financial inclusion has emerged as a primary focus for governments worldwide (Sun & Wang, 2023). Consequently, there has been a surge in statistical research aimed at addressing this imperative. The abundance of information on this issue has facilitated the development of various methodologies; many of these operate under the premise that accurate measurement is essential for assessing the efficacy of public policies aimed at improvement. However, it is worth noting that the type of information gathered through surveys may have varying potential in different scenarios.

On this note, the causality behind an individual's consideration for financial inclusion has predominantly been approached from a theoretical and qualitative standpoint. Many discussions on enhancing financial inclusion primarily focus on customer preferences and income levels. However, given the multidimensional nature of this issue, it is important to consider other factors or features that may influence individuals' decisions regarding whether to access the financial system or not.

The present document outlines a machine learning procedure conducted on the Encuesta Nacional de Inclusión Financiera (ENIF, or National Financial Inclusion Survey) (INEGI, 2021) dataset, focusing on financial inclusion in Mexico during

Itzel Coquis Rioja, Mario Iván Contreras Valdez
Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case
Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

2021. The aim is to characterize individuals falling within the broader definition of financial inclusion. By identifying the a priori conditions for inclusion in this dataset, policymakers can devise more effective public policies to enhance the likelihood of individuals belonging to the financial system. This research endeavors to transition from qualitative and theoretical perspectives to statistical and quantitative methodologies; through these, the impact of different variables can be evaluated to incentivize financial inclusion effectively. The importance of this research lies in its utilization of classification artificial intelligence models. These models yield two crucial outcomes: firstly, they identify the variables that contribute to the conditional probability of individuals being considered in financial inclusion. Secondly, they uncover nonlinear relationships between features and the response variable. Furthermore, the results of this research can be utilized by governments to implement public policies aimed at increasing the likelihood of individuals accessing the financial system.

To structure the present paper, Section 2 develops the literature review; Section 3 presents the data and methodology to evaluate the relationship; Section 4 presents the main results of these models and develops the interpretation and analysis of them; finally, Section 5 sets out the conclusions, and the last section includes the references.

2. Literature Review

Financial inclusion is a term that can be traced back several decades. The Community Reinvestment Act (1977) was enacted by the US to regulate banks, preventing them from focusing only on rich districts and obliging them to provide financial services, regardless of the population's income level. In this regard, Leyshon and Thrift (1995) understand financial exclusion to be all those mechanisms that serve to prevent certain social groups from entering the financial system through the availability of financial instruments. Further refinements to the definition encompass the specification that it prevents poor or unprotected groups from gaining access to the formal financial system or to regulated financial instruments in a low-cost and safe environment, where credit might be the instrument most required (Conroy, 2005; Mohan, 2006; Rangarajan Committee, 2008; Marín & Schwabe, 2018).

Its importance in incentivizing other economic variables is closely connected with financial inclusion. One of the earliest works in this regard was done by the economist

Schumpeter (1911), in a text where the author shows that the financial sector has the potential to boost economic growth and economic development. The primary function by which it operates is the ability of the banking system to facilitate capital accumulation, thereby enabling firms to utilize the capital in the production process. Consequently, this leads to an increase in the output and productivity of productive factors, ultimately enhancing the purchasing power of households. (King & Levine, 1993; Beck et al., 2007). However, those initial analyses only addressed innovation within the financial system, which is also known as financial development. In this regard, Johnson and Arnold (2014) note that public policy has undergone a shift in its priorities, from Financial Development to Financial Inclusion. In this case, focusing efforts not only on the strength of the banking system and the free flow of capital is a first step, but the goal nowadays has also turned towards enabling the economically vulnerable population to have access to this financial system.

In the case of Mexico, Segovia and Cepeda (2024) use panel data to measure the level of financial inclusion in different states from the periods of 2005 to 2018. Their findings show that the increase in banking credit for non-financial firms has the potential to increase GDP per capita. In this instance, the Region variable seems to be a proper indicator for capturing the effect described in this document. In this same vein, De la Cruz and Alcántara (2011) provide statistical evidence for the positive relationship between the development of the private banking system and economic development. A particular experience of this relationship is described in Bruhn and Love (2014), where the incursion of Banco Azteca in the banking system via its Elektra stores shows a shift in poverty reduction because of the incorporation of financial and commercial infrastructure, making clear the importance of this element in the financial inclusion of Mexican society.

The growing importance of financial inclusion has generated another issue in relation to the definition and subsequent measurement of the variable. Huong et al. (2023) expose this problem as a global phenomenon, where the debate is centered on a universally accepted term; however, the shared understanding that can be reached is that financial inclusion entails access to formal financial services at an affordable cost. In this regard, the authors also note that currently there is a lack of an adequate method for assessing financial inclusion. The main problem arises from the fact that proposed indexes have arbitrary weights for the characteristics of people who are considered to have financial inclusion.

In view of this, Sarma (2015) developed a financial inclusion index for 81 countries using a number of economic indicators that worked as a reference point; further

Itzel Coquis Rioja, Mario Iván Contreras Valdez
Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case
Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

research has been extended using the same procedure for 98 more countries (Arora, 2014). For a local case, Dircio-Palacios Macedo et al. (2023) proposed the construction of an index for the municipalities of Mexico. Their initial results indicate that there are variations among municipalities, making it challenging to implement a universal index.

Another perspective on this same logic refers to the concept of digital financial inclusion (DFI). Xi and Wang (2023) provide a number of outcomes concerning the influence of DFI on economic growth. Their conclusions highlight certain attributes that could serve as a more concise definition since technological applications might expose unsuspecting individuals to scams and fraud in electronic mediums (Yue et al., 2022). Because of this, Schuetz and Venkatesh (2020) argue that one of the main obstacles to reaching financial inclusion, along with geographical access and the high relative cost of certain services, is the presence of financial illiteracy. All of these sources remark on the need not only for infrastructure or technology but also for the understanding of personal finance.

So far, the impacts of financial inclusion and some of the problems in measuring them can be analyzed; however, from a complete perspective, the factors contributing to people's involvement should also be considered. In that respect, Cruz-García et al. (2020) developed a model to identify the probability of a given municipality being considered to offer financial inclusion. Their results show that geography might play a major role in the determination of financial inclusion, as certain statistics seem to have proven their ability to predict the probability. Jamil et al. (2024) present a complementary analysis of the impact that age has on this topic. Their hypothesis starts from the fact that the older population who have retired from the labor market might need to have a strong understanding of their financial situation in order to live mostly from their savings. Therefore, this characteristic within the population could serve as an indicator of whether people have knowledge about this topic or not.

3. Data and Methodology

The present study uses the Encuesta Nacional de Inclusión Financiera (ENIF) (INEGI, 2021), which has been conducted in Mexico every three years since 2012, when the Instituto Nacional de Estadística y Geografía (INEGI, Statistics and Geography National Institute) and the Comisión Nacional Bancaria y de Valores (CNBV, National Baking and Stock Commission) joined forces to obtain relevant

statistics on financial inclusion. The purpose of these exercises is to identify the main challenges to financial inclusion in the country; on this basis, public policies that are the responsibility of the Consejo Nacional de Inclusión Financiera (CONAIF, National Council for Financial Inclusion) could be implemented in future proposals.

The current data comes from the latest survey conducted in 2021, with information compiled from June 28 to August 13. Using a confidence level of 90%, the sample was defined to be local and nationally representative. The sample size consisted of 13,554 people, representing 90,328,320 people over 18 years of age. The questionnaire was separated and compiled into four modules with different aggregate levels of the households. The current study is developed based on the TModulo file, which contains the answers from the individual questionnaires completed in the survey. The data consists of a rectangular grid of 13,554 rows and 382 columns with the variables and identifiers.

To conduct causality modeling, two artificial intelligence models are introduced to analyze the categorical variable indicating whether an individual demonstrates financial inclusion or not. The aim of this approach is to discern the directional impact of individual features that influence the likelihood of belonging to the financial system. The outcomes of these models will enable the identification of exogenous variables that drive individuals to be part of the financial system. Such insights can then be leveraged by policymakers to enhance financial inclusion indicators.

The first model we propose is logistic regression with categorical exogenous variables. Originally introduced by Berkson (1944) to investigate the impact of certain drugs on patient survival, logistic regression has evolved to become a cornerstone of artificial intelligence models. It serves as a bridge between black-box classification models and traditional inference statistics. The logistic function, which is central to this model, can be analytically expressed as follows:

$$F(z_i) = \frac{e^{z_i}}{1 + e^{z_i}} = \frac{1}{1 + e^{-z_i}}$$

Where e is the exponential number and F represents the cumulative logistic function. In sum, using this model allows us to compute the probability P_i for the observation $Y_i = 1$ where the value of Z_i represents a linear function; that is:

$$P_i = \frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}}}$$

Itzel Coquis Rioja, Mario Iván Contreras Valdez
Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case
Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

Moreover, since some variables in the survey are categorical, with responses being either Boolean or categorical, we incorporate transformations into dummy variables. This allows for comparisons with individual responses from the questionnaire. Conceptually, the model constructed resembles parallel lines containing several categories in the result. In consequence, the estimates derived from logistic regression can be interpreted as the probability differences between individual features.

To further enhance this analysis, a tree-based model is proposed to complement the causality analysis. Such models aim to segment the response space of observations, enabling a discriminant function to allocate observations based on their features. In this context, the decision tree will be a classification model, given that the predictor consists of dummy variables.

The method for adjusting such a tree involves binary splitting with the features of the observation. Each time a bifurcation occurs, a node is created with two branches that can further expand using other variables. Since the objective of the model is to maximize the probability of the observation belonging to a specific region or class, the setup is defined as follows:

$$E = 1 - \max_{k}(\hat{p}_{mk})$$

Where $\hat{p}_{mk}$ is the proportion of observations in the m region belonging to the k class. To make the empirical fit, however, another measure is preferred, for tree-based models the Gini Index can be defined as:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

This expression allows for values of the Bernoulli Density Function to operate in a mean approach to maximize the likelihood of the observation belonging to a category or not.

The rationale behind selecting the previous models over more advanced machine learning ones is their interpretability. In many classification problems tackled with artificial intelligence, the primary objective is to achieve the best possible prediction for the data. However, this often results in the creation of black-box models, where the features or exogenous variables are not directly interpretable. While these models serve as powerful forecasting tools, they lack transparency in understanding the underlying behavior of the phenomenon.

The logistic regression and decision tree models belong to a subset of machine learning models that accommodate non-linear data behavior, in contrast to traditional multivariate linear models. However, they retain the necessary ability to explain and understand the relationship between the response and exogenous variables. Using more complex and flexible models may lead to difficulties in measuring the direction and impact of variables, as traditional linear models are unable to capture non-linear impacts (James et al., 2017). Additionally, compared to Probit models, logistic regression allows for the use of a distribution with heavier tails than the Gaussian Density. Therefore, in the presence of possible outliers, logistic regression proves to be more robust (Gujarati & Porter, 2009).

As mentioned earlier, the dataset comprises all observations from the survey. To assess the a priori variables that may contribute to financial inclusion, the selected variables mainly pertain to household financial inclusion characteristics. The following variables were chosen to identify the features influencing inclusion in the financial system. Table 1 presents the variable names along with brief descriptions of their meanings and categorical levels (see Table 1).

Table 1. Variable Description for Exogenous Features

Variable	Description
Location size	Number of inhabitants of the location; it consists of four categories "More than 100,000," "Between 15,000 and 99,999," "Between 2,500 and 14,999" and "Less than 2,500."
Region	Six regions in which INEGI divides the territory; "Northwest," "Northeast," "West," "CDMX," "Southwest" and "Southeast."
Sex	Sex: "Male" or "Female."
Age	Age, with valid values in the range of 18 to 96.
Education	Completed schooling level, ranging from "None" to "Postgraduate."
Civil	Marital status: "Single," "Free Union," "Separated," "Divorced," "Widow," "Married."
Indigenous	Proxy variable to characterize if the person belongs to an indigenous group; the question only asks if the person speaks an indigenous language.
Government Program	If he or she is a beneficiary of a government welfare program.

Itzel Coquis Rioja, Mario Iván Contreras Valdez
Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case
Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

Variable	Description
Occupation	Main occupation: "worker," "unpaid worker," "unemployed," "student," "homemaker," "retired," "disabled" and "not economically active."
Labor Class	"Worker," "unpaid worker," "laborer," "self-employed," "employer."
Income	Monetary amount of income measured per month.
Income Frequency	Frequency of payment: "weekly," "fortnight," "monthly" and "yearly."
Stable	If the perceived income is regular in frequency or not.
Smartphone	If the person owns a smartphone.
Budget	If the person has a budget control over its monthly expenditure.
Annotations	If the person uses a system to keep track of his or her own expenditure.
App	If the person uses a digital tool to keep control over their expenditure.
Resources	If the person, in any of the last three years, experienced a shortage of money by the end of the month.
Considerations	How often does the person consider if there is enough money to make a purchase: "always," "sometimes," "never," "not answered," "don't know."
Future	How often does the person prefer to consume rather than save for the future: "always," "sometimes," "never," "not answered," "don't know."
Goals	How often the person set a financial or economic goal and work to achieve it: "always," "sometimes," "never," "not answered," "don't know."
Driver	How often does income and expenditure take control over the life of the person: "always," "sometimes," "never," "not answered," "don't know."
Present	How much does the person agree with the statement: they prefer thinking about the present rather than the future: "agree," "neither agree nor disagree," "disagree" "not answered," "don't know."
No Income	How long would the person be able to maintain their current level of expenditure: "don't know," "not answered," "no savings," "1-3 weeks," "1-3 months," "3-6 months" and "6+ months."
Savings	In the last 12 months, how much money was the person capable of saving in reference to their income: "1 week," "1 fortnight," "1 month," "1+ month," and "don't know."
Financial Education	If the person has ever taken a course in financial education.

Source: Prepared by the authors using the variable descriptor of ENIF survey referring to the module data subset.

From another research perspective, it is imperative to establish the definition of financial inclusion. According to INEGI (2021), it entails "the access to and use of formal financial services under proper regulation that guarantees consumer protection schemes and promotes financial education." While this definition encapsulates the multidimensional nature of financial inclusion, it is essential to devise a measurable way to assess it. In this study, following the logic of the ENIF design, we have chosen the so-called "filter" variables to represent the various categories of financial inclusion. Three levels or questions have been considered, and their interpretations as variables are presented in Table 2 (see Table 2).

Table 2. Variable Description for Predicted Variable

Variable	Description
Savings	Answer to the question: Did you save any money in a financial account?
Payroll Card	Answer to the question: Do you have a payroll card?
Debit Card	Answer to the question: Do you have a debit card?
Inclusion	Boolean variable if all previous answers are "Yes." This will be used as the dependent variable in the classification

Source: Prepared by the authors using the variable descriptor of ENIF survey referring to the Module data subset.

To conduct this study, two models with the dependent variable in Table 2 will be conducted separately. The variables in Table 1 will be used as features in the machine learning models previously presented. To perform the evaluation, the confusion matrix approach will be implemented. In this sense, 60% of the persons will be randomly selected as part of the training dataset, and the remaining 40% will correspond to the testing dataset. The main statistics to evaluate are the accuracy ratio, McNemar's test, and Cohen's Kappa. Furthermore, the interpretability of the models will allow the evaluation of the significance and impact of the variables included to obtain the best model to predict the presence, or not, of financial inclusion.

3. Results

To implement the previous models in the datasets we begin by gathering and transforming the variables into dummy ones. By doing so, the model can be executed as a parallel line's Logit model and decision tree. At the same time, a balanced data

Itzel Coquis Rioja, Mario Iván Contreras Valdez
Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case
Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

transformation is proposed as the data contains different amounts of observations in the classifications that were previously defined. The first step is to identify the number of persons belonging to each side of the financial inclusion, and then a random sampling of the same length is taken for both sides. This provides balanced data that can be analyzed with the prior variables without having a size skewness in the over or under-identification of the features. Table 3 contains the counting for each of the classifications in the variable presented in Table 2 (see Table 3).

Table 3. Counting Statistics for Response Variables

Variable	Yes	No
Inclusion	1632	11922

Source: Prepared by the authors using the variable descriptor of ENIF survey referring to the Module data subset.

In such case, the number of observations per model will be the value for the "Yes" response. The remaining observations will be randomly sampled to fit the same value. A backward-looking selection algorithm will be implemented to present only the best models in the model selection for the main features to be included.

The first model to consider is the one with the "inclusion" variable in the Logit specification. Table 4 contains the results for the regression estimates and p-value for t-test significance values (see Table 4) and Table 5 presents the confusion matrix with the main statistics to conduct the evaluation (see Table 5). It is important to mention that in the model selection process, the multicollinearity phenomena were controlled using the variance inflation factor. The presented model was the best in calibration.

Table 4. Logit Results for Inclusion as Response Variable

Coefficient	Estimate	Standard Error	p-value
Intercept	-5.47	1.55	4.27×10^{-4}
Income	1.14×10^{-4}	2.22×10^{-5}	2.52×10^{-7}
Education (Elementary)	4.43×10^{-1}	1.32	7.38×10^{-1}
Education (Middle School)	1.38	1.30	2.88×10^{-1}
Education (Middle Technical)	1.89	1.45	1.95×10^{-1}
Education (High School)	2.12	1.30	1.04×10^{-1}

Coefficient	Estimate	Standard Error	p-value
Education (High Technical)	3.45	1.43	1.58×10^{-5}
Education (Bachelor)	2.90	1.31	2.69×10^{-2}
Education (Post Graduate)	4.84	1.69	4.12×10^{-3}
Government Program (Yes)	1.01	5.17×10^{-1}	4.97×10^{-2}
Smartphone (Yes)	1.15	3.20×10^{-1}	3.34×10^{-4}
Resources (Not Enough money)	-5.25×10^{-1}	1.92×10^{-1}	6.18×10^{-3}
No Income (Not answered)	5.50×10^{-1}	1.29	6.71×10^{-1}
No Income (No savings)	-3.14×10^{-2}	7.69×10^{-1}	9.67×10^{-1}
No Income (1-3 weeks)	4.33×10^{-1}	7.68×10^{-1}	5.73×10^{-1}
No Income (1-3 months)	1.29	7.57×10^{-1}	8.74
No Income (3-6 months)	1.56	7.81×10^{-1}	4.56×10^{-2}
No Income (6+ months)	2.45	7.95×10^{-1}	2.03×10^{-3}
Labour Stability (Fixed)	1.27	1.88×10^{-1}	1.15×10^{-11}
Location Size (100000+)	6.48×10^{-1}	2.84×10^{-1}	2.23×10^{-2}
Location Size (15M - 99M)	5.36×10^{-1}	3.31×10^{-1}	1.06×10^{-1}
Location Size (2.5M - 15M)	6.23×10^{-1}	3.57×10^{-1}	8.14×10^{-2}
Financial Education (Yes)	1.26	3.93×10^{-1}	1.35×10^{-3}

Source: Prepared by the authors with information from ENIF. The results are displayed as the maximum likelihood estimator for the coefficients. The p-value refers to the significance t-Test.

Table 5. Confusion Matrix for Inclusion as Response Variable

	No	Yes
No	256	61
Yes	66	361
Accuracy	0.83	
Kappa	0.65	
p-value McNemar's Test	0.72	

Source: Prepared by the authors. The confusion matrix in the first part of the table represents the empirical category the observation belongs to in the rows, while the columns are for the forecast. The main diagonal counts for the observations are correctly identified, while the secondary diagonal is badly allocated. The accuracy ratio is the proportion of correctly identified observations relative to the whole sample. Kappa stated for a cleaner accuracy interpretation, while McNemar's Test captures the bias error, the null hypothesis states that the incorrectly identified observations are the same.

Itzel Coquis Rioja, Mario Iván Contreras Valdez
Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case
Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

The original model included all the previously mentioned variables; however, the significance test and evaluation using the confusion matrix did not provide evidence of their significance over the response variable. Additionally, we considered variables that may cause multicollinearity, and in this instance, the VIF criterion was employed to refine the model, retaining only the variables without severe multicollinearity. Therefore, only the best model is presented here. In this case, considerations were made to assume significance under a 95% confidence level, although in some cases, we also allowed for a 90% confidence level if the interpretation of such a variable aligned consistently with results obtained by previous authors. The selection criteria also considered the highest value for Accuracy and Cohen's Kappa, along with the non-rejection of the McNemar's Test of bias in the error.

To interpret the model, each variable will be explained based on the increment or decrement of the probability of having financial inclusion. It is worth mentioning that the estimated values presented in Table 4 correspond to the log-odds of the model, so only the sign of the estimate can be read as an increment or decrement; still, to have a proper reading, it is necessary to transform such values into traditional probabilities.

The first variable to be significant in the model is the observation's monthly income, t a numeric variable measured in Mexican pesos, so the marginal increment in probability is 0.0025% per extra monthly peso of income. This can be understood as a positive relationship between income and the likelihood of having this type of financial inclusion.

The second variable incorporated in the model is Education. In this case, we have a categorical variable with eight levels, so the estimates are compared with respect to the lowest level of education which, in this case, is labeled as "None." In essence, we are presenting the results for dummy variables in the education category. The first elementary and middle school levels are not statistically significant compared to those with no education at all, meaning that there is no statistical difference between those three groups. However, from the level of middle technical, high school, bachelor, and postgraduate, a significant alteration can be perceived. Compared to those without any education, middle technical individuals have an increase in probability of being in financial inclusion of 25.82%, the high school increments by 35.13% the chances, high technical tends to have a marginal increment of 32.42%, bachelors have a difference of 46.94% and postgraduates of 36.37%. As may be appreciated, the increase in financial inclusion is significant when we isolate this category as a comparison variable. The interpretation of this result can be explained

by the capabilities of people reaching these educational levels; as their income tends to be higher and more regular than those in jobs with lower education levels, the incentive and ability to save money for the future increases as well. Another interesting aspect to contemplate is the fact that in those educational levels, extra information in financial education is provided, allowing individuals to be conscious about such topics.

The next variable is being the beneficiary of any government program. In this sense, the Mexican Government has developed several social programs intended to aid those with a higher economic vulnerability. In doing so, many programs lean toward the digital transfer of resources. The result is that the population in such conditions was incentivized to enter the financial system in order to benefit from these programs. When isolating for the marginal impact of this variable compared to those without government aid, the probability increases by 17.9%, allowing for a positive relationship.

The Boolean variable of owning a smartphone follows. This technology enables various activities in daily life, one of which is accessing financial services or applications to understand personal finance better. According to WallStreetZen (2023), gen Z individuals learn about money using YouTube and TikTok digital platforms. Although this can potentially enhance financial inclusion, the survey also mentions the incentives for content creators to provide false information to generate more engagement. In this sense, it is worth noticing and questioning not only the quantity but the quality of financial inclusion, as greater access to financial services also has the potential to be used to scam or to provide incorrect information to financial consumers. In this respect, possessing a smartphone provides an extra 27.14% probability of attaining financial inclusion.

Next comes a variable about the income and expenditure level of the individual. The question refers to whether or not the person arrives at the end of the month with enough money. The reference category is that the person has "enough money" by the end of the month. For the estimated value of the coefficient, the difference is significant and has an impact on the probability of -11.58%. The previous result explains that a person with insufficient money is less likely to have financial inclusion. This aspect, which might seem trivial, is not so in an economic context where households present high expenditure levels that are not met by their current income. This inability to save money and fear to ask for consumer credit might explain some of the low levels of financial inclusion in Mexico. In this case, creating instruments suited for those types of individuals, either in saving

Itzel Coquis Rioja, Mario Iván Contreras Valdez
Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case
Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

or allowing for credit, may be more productive than simply expanding the current infrastructure.

The next variable to be incorporated into the model delves into another aspect of purchasing power and future preparedness. It aims to determine how long an individual could sustain their current expenses if they suddenly ceased to receive income. The reference category for this question is "Don't Know." The results indicate that there is no statistically significant difference among the responses "No Savings," "Not Answered," and "1-3 weeks." However, the real impact surfaces among those capable of maintaining their expenses for one month or more. For individuals able to sustain their lifestyle for "1-3 months," the probability of financial inclusion increases by 24.96%. This likelihood rises to 26.1% for those able to manage for "3–6 months," and for those who can sustain themselves for more than six months, the chances of financial inclusion increase by a substantial 35.85%. This underscores the importance of the income-expenditure relationship, highlighting the need for financial instruments that facilitate savings and enhance financial security.

Another variable that was shown to be statistically significant was the stability of income perception. People with a fixed income present an increase in probability compared to variable income of 27.3%. In this case, the Mexican labor structure works against financial inclusion as a considerable portion of the labor market belongs to the informal sector. In this instance, the inability to access to a financial instrument may have caused an exacerbated impact during the pandemic because of the contingency and reduction in consumption.

The location size may be used as a proxy variable for the infrastructure availability of the individual. In Mexico, INEGI measures this variable as "less than 2500," "between 2500 and 15,000," "between 15,000 and 99,000" and "more than 100,000" people. The first two categories can be described as rural areas, while the latter are urban or semi-urban locations. For the interpretation of this estimate, the reference point is for communities with fewer than 2500. Therefore, when the individual belongs to a community between 2500 and 15,000, the chances to be part of financial inclusion increases by 12.27%; when it has between in 15,000 and 99,000 persons, the probability increases by 10.82%. Despite that, it is worth mentioning that these results are valid only under a 90% confidence level, so it might be argued that there is no difference in these three categories. However, in the urban area where the population reaches more than 100,000 people, the probability increases by 14.27% with a confidence level of 95%. As such, the result of this variable can be interpreted as a positive impact of infrastructure on financial inclusion.

261

The final variable to be included is the response to the question of whether the person has ever taken a "personal finance" course. As such, the increment in probability for financial inclusion due to this fact is 22.1%. This strengthens the notion that having access to this type of information on the benefits and types of financial instruments, as well as how to make best use of them, has a positive and significant impact.

For the second model, the results of the estimation are better presented graphically. Figure 1 presents the branches and conditions for the identification of the observation belonging to the financial inclusion (1) or not (0) (see Figure 1). Table 6 presents the confusion matrix for the decision tree (see Table 6).

Figure 1. Tree Identifying Financial Inclusion
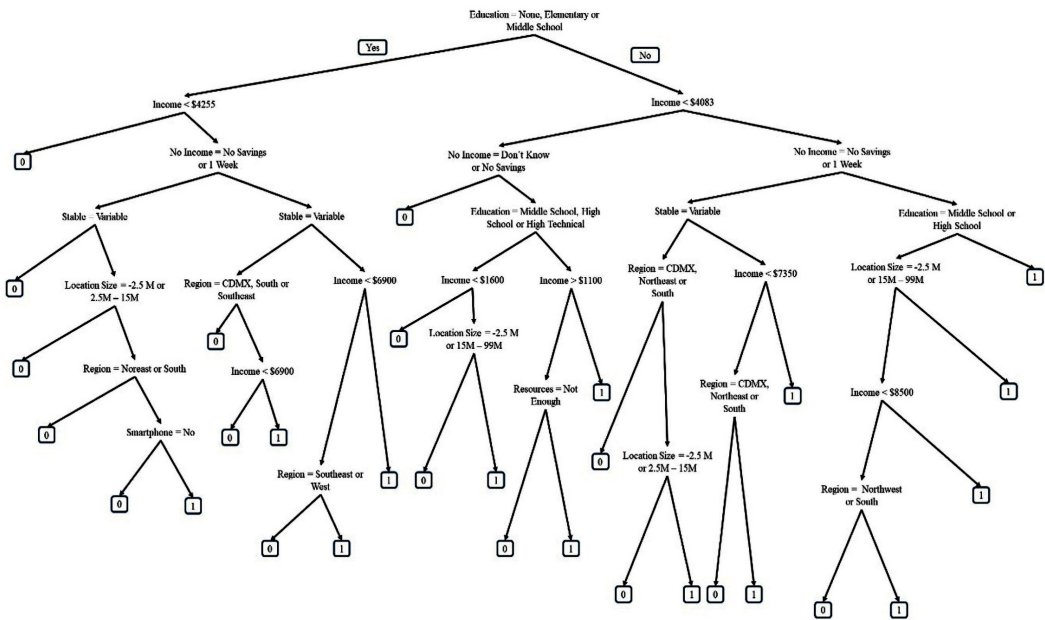
Source: Prepared by the authors to represent the variables as nodes in the decision tree. The left branches are for when the condition is true and the right branches when the statement is false. The 0 represents the observation without financial inclusion and the 1 represents those who have it.

Itzel Coquis Rioja, Mario Iván Contreras Valdez
Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case
Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

Table 6. Confusion Matrix for Decision Tree Model

	No	**Yes**
No	245	69
Yes	79	348
Accuracy	0.80	
Kappa	0.59	
p-value McNemar's Test	0.12	

Source: Prepared by the authors. The confusion matrix in the first part of the table represents the empirical category the individual belongs to, in the rows, while the columns are for the forecast. The main diagonal counts for the individuals correctly identified while the secondary diagonal those badly allocated. The accuracy ratio is the proportion of correctly identified observations relative to the whole sample. Kappa stated for a cleaner interpretation of the accuracy while McNemar's Test captures the bias error; the null-hypothesis states that the incorrectly identified observations are the same.

For the interpretation of this tree, the paths leading to the observation belonging to the financial inclusion group were gathered in Table 7 (see Table 7). As a result, 14 sub-groups were identified by the algorithm as the combination of features that allow a person to belong to the financial inclusion group. In each row the description of the characteristics is included.

Table 7. Features Identified by the Decision Tree that Allows a Person to Belong to the Financial Inclusion Category

Sub-group	Characteristics
1	No education, elementary or middle school Income > $4255 No savings or for less than a week Stable income Location size larger than 15,000 Not from northeast or south region Owns smartphone
2	No education, elementary or middle school Income > $6900 Savings for more than a week Variable Income From northeast or northwest region

Sub-group	Characteristics
3	No education, elementary or middle school Income between $4,255 and $6,900 Savings for more than a week Fixed income Not from southeast or northwest region
4	No education, elementary or middle school Income > $6,900 Savings for more than a week Fixed income
5	Middle school, high school or high technical Income < $1,083 Savings for at least 1 week Location with more than 100,000 people
6	Bachelor or postgraduate education Income > $1,083 Savings for at least 1 week Enough resources to reach the end of the month
7	Bachelor or postgraduate education Income < $1,083 Savings for at least 1 week
8	Education above middle school Income > $4,083 No savings or less than a week Variable income Not from CDMX, northeast or south Location size above 100,000 people
9	Education above middle school Income between $4,083 and $7,350 No savings or less than a week Fixed income Not from CDMX, northeast or south
10	Education above middle school Income > $7,350 No savings or less than a week Fixed income
11	Education above High School Income > $4,083 Savings for more than a week

Itzel Coquis Rioja, Mario Iván Contreras Valdez
Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case
Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

Sub-group	Characteristics
12	Middle or high school Income > $4,083 Savings for more than a week Location with more than 100,000 people
13	Middle or high school Income > $8,500 Savings for more than a week Location with more than 100,000 people
14	Middle or high school Income between $4,083 and $8,500 Savings for more than a week Location with more than 100,000 people Not from northwest nor south

Source: prepared by the authors using information from the decision tree model.

It is noteworthy that the most influential variables in identifying a person's level of financial inclusion once again center around education levels, with middle and high school education serving as pivotal benchmarks that define the disparity. Following closely is the individual's region, where sectors in the north exhibit a higher propensity to utilize financial instruments. Additionally, the size of the location plays a significant role in differentiating observations, enabling individuals with lower economic standing to access financial instruments, particularly in areas with larger population sizes. When interpreted as a proxy variable for infrastructure, this underscores the critical importance of such features in fostering financial inclusion.

When comparing the two models, it's important to highlight that those who opted to incorporate all the previously mentioned variables, driven by a principle of simplicity and effectiveness in achieving predictability, refined the models to showcase the most optimal results attainable. In both cases, some variables are incorporated and represent a major influence on the discriminatory analysis. For instance, the education level, location size, income, and income stability are present. On the other hand, the Logit model includes, explicitly and with a huge impact in probability, the feature of the individual having any type of personal financial education. The second important different variable is the access to a smartphone, which the Logit model uses as a positive relationship with financial inclusion; however, the decision tree does not use it.

Concerning the results in the machine learning evaluation, the accuracy level of both models is around 80%; however, logistic regressions seem to have a better performance when using the Kappa criteria. Also, both algorithms present an adequate value for the McNemar skewness test, allowing for a proper fit and unbiased errors. It can be resolved that artificial intelligence algorithms under the classification basis are a proper way to make the study of financial inclusion drivers. Furthermore, the main advantage of these models is their capability to detect the non-linear effect of independent variables over the response without having to sacrifice the interpretability of the results, as it may happen with more sophisticated models like neural networks or support vector machines.

In this instance and using the Kappa and accuracy criteria, the Logit models seem to have a better performance in the detection of the variables belonging to financial inclusion. Nevertheless, the decision trees also work as a reinforcement of the importance of variables in the features that drive the financial inclusion phenomena.

4. Conclusions

The current document highlights the significance of financial inclusion as an objective for any developing nation. The international attention this variable is receiving allows for multiple studies to be performed and a plethora of causes and effects to be proposed. To contribute to this, a machine learning approach is suggested. As part of this, Logit and decision Tree models are used as a technological tool that makes it possible to capture non-linear relationships for features of the observations and the probability of belonging to the financial inclusion sector. By using analysis related to statistical and artificial intelligence evaluation, the best models are computed with information from the ENIF survey. The main contribution of this paper is to quantify the impact of some characteristics of the observations on the probability of a person having financial inclusion. As such, it provides a better understanding of the phenomena that can be implemented in public policies to improve the indicator.

The main findings of the models provide information about the importance of education levels over financial inclusion. Individuals reaching middle school have a higher and significant improvement in the probability of being part of the financial system. This situation can be attributed to education programs where topics on personal finance are introduced too late in the curriculum. Consequently, individuals who only attain a basic level of education may never receive formal instruction on

Itzel Coquis Rioja, Mario Iván Contreras Valdez
Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case
Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

financial matters. To control for this condition, the explicit question about receiving a course in personal finance is included. With it, it can be noticed that having this extra knowledge is a significant driver to increase the probability.

Education is one of the main studied and supported variables to be included in the financial inclusion documents; however, the other features that have a major impact on it are the amount of wealth, the consistency of a perceived income, and the size of the location. In terms of the amount of wealth, the model found that the marginal effect of having enough money to survive for at least one month without any income is almost equal to the educational level.

▬ References

Arora, R. U. (2014). Access to finance: An empirical analysis. *The European Journal of Development Research*, *26*(5), 798–814. https://doi.org/10.1057/ejdr.2013.50

Beck, T., Demirguc-Kunt A., & Martinez Peria, M. S. (2007). Reaching out: Access to and use of banking services across countries. *Journal of Financial Economics*, *85*(1), 234–266. https://doi.org/10.1016/j.jfineco.2006.07.002

Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, *39*(227). https://doi.org/10.2307/2280041

Bruhn, M. & Love, I. (2014). The real impact of improved access to finance: Evidence from Mexico. *The Journal of Finance*, *69*(3), 1347–1376. https://doi.org/10.1111/jofi.12091

Community Reinvestment Act (1977). Federal Reserve.

Conroy, J. (2005). APEC and financial exclusion: missed opportunities for collective action? *Asia Pacific Development Journal*, *12*(1), 53–79.

Cruz-García, P., Dircio-Palacios Macedo, M. del C. & Tortosa-Aucina, E. (2020). Financial inclusion and exclusion across Mexican municipalities. *Regional Science Policy & Practice*, *13*(5), 1496–1526. https://doi.org/10.1111/rsp3.12388

De la Cruz Gallegos, J. L. & Alcántara Lizárraga, J.A. (2011). Crecimiento económico y el crédito bancario: un análisis de causalidad para México . *Revista de Economía*, *28*(77), 9–38.

Dircio-Palacios Macedo, M. del C., Cruz-García, P., Hernández-Trillo, F. & Tortosa-Ausina, E. (2023). Constructing a financial inclusion index for Mexican municipalities. *Finance Research Letters*, *52*(C), article 103368. https://doi.org/10.1016/j.frl.2022.103368

Demirguc-Kunt, A. & Klapper, L. (2012). Measuring financial inclusion: The global findex database (No. 6025). The World Bank.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Application in R*. Springer.

Jamil, A. R. M., Law, S.H., Khair-Afham, M.S. & Trinugroho, I. (2024). Financial inclusion and income inequality in developing countries: The role of aging populations. *Research in International Business and Finance*, *67*(PA). https://doi.org/10.1016/j.ribaf.2023.1

Gujarati, D. & Porter, D. (2009). *Basic Econometrics*. McGraw-Hill Series Economics.

Huong Tram, Thi Xuan, Lai, Tien Dinh, Huong Nguyen, Thi Truc. (2023). Constructing a composite financial inclusion index for developing economies. *The Quarterly*

Itzel Coquis Rioja, Mario Iván Contreras Valdez
Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case
Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

Review of Economics and Finance, 87 (257–265). https://doi.org/10.1016/j.qref.2021.01.003

Instituto Nacional de Estadística Geografía e Informática (INEGI). (2021). Encuesta Nacional de Inclusión Financiera (ENIF) 2021.

Johnson, S. & Arnold, S. (2014) Inclusive financial markets: is transformation under way in Kenya? *Development Policy Review*, 32 (5), 639–642. https://doi.org/10.1111/j.1467-7679.2012.00596.x

King, R. G. & Levine, R. (1993) Finance and growth: Schumpeter might be right. *The Quarterly Journal of Economics, 108*(3), 713–737. https://doi.org/10.2307/2118406

Leyshon, A. & Thrift, N. (1995). Geographies of financial exclusion: financial abandonment in Britain and the United States. *Transactions of the Institute of British Geographers*, *New Serie*s *20*(3), 312–341. https://doi.org/10.2307/622654

Marín, A. G. & Schwabe, R. (2018). Bank competition and financial inclusion: Evidence from Mexico. *Review of Industrial Organization*, *55*(2), 257–285. https://doi.org/10.1007/s11151-018-9673-5

Mohan, R. (2006). Economic growth, financial deepening, and financial inclusion. Address at the Annual Bankers' Conference 2006, Hyderabad on November 3, 2006. http:// rbidocs.rbi.org.in/rdocs/Speeches/PDFs/73697.pdf

Rangarajan Committee. (2008). Report of the Committee on Financial Inclusion. Government of India.

Sarma, M. (2015). Measuring Financial Inclusion. *Economic Bulletin*. *35*(1), 604–611.

Schuetz, S. & Venkatesh, V. (2020). Blockchain, adoption, and financial inclusion in India: Research opportunities. *International Journal of Information Management*, *52*. https://doi.org/10.1016/j.ijinfomgt.2019.04.009

Schumpeter, J. (1911). *The theory of economic development*. Harvard University Press.

Segovia, M. A. F. & Cepeda, L. E. T. (2024). Financial development and economic growth: New evidence from Mexican States. *Regional Science Policy & Practice*, article 100028. https://doi.org/10.1016/j.rspp.2024.100028

Sun, T. & Wang, X. (2023). Adoption of financial inclusion in a world of depleting natural resources: the importance of information and communication technology in emerging economies. *Resources Policy,* 85. https://doi.org/10.1016/j.resourpol.2023.103901

Tram, T. X. H., Lai, T. D. & Nguyen, T. T. H. (2023). Constructing a composite financial inclusion index for developing economies. *Quarterly Review of Economics and Finance*, *87*, 257–265.https://doi.org/10.1016/j.qref.2021.01.003

Varian, Hal R. (1992). *Microeconomic Analysis*. Norton.

WallStreetZen. (2023). Where Did Gen Z Learn About Money. https://www.wallstreet-zen.com/blog/genz-money-social-media-survey/

Xi, W. & Wang, Y. (2023). Digital financial inclusion and quality of economic growth. *Heliyon, 9*. https://doi.org/10.1016/j.heliyon.2023.e19731

Yue, P., Korkmaz, A.G., Yin, Z. & Zhou, H. (2022). The rise of digital finance: financial inclusion or debt trap? *Finance Research Letters,* 47(A). https://doi.org/10.1016/j.frl.2021.102604

Itzel Coquis Rioja, Mario Iván Contreras Valdez
Causality Study on Financial Inclusion Issues with Data Science Techniques: The Mexican Case
Estudio de causalidad sobre problemas de inclusión financiera con técnicas de ciencia de datos: el caso de México

▬ About the authors

Itzel Coquis Rioja graduated from Tecnológico de Monterrey (ITESM) Campus Mexico City with a degree in Business Administration and holds a Master's in Energy Administration and Renewable Resources. She is currently a doctoral student in Financial Sciences. After completing her studies, she began her professional career at Coca-Cola Femsa, in Human Resources, in 2011. She worked in the area of market intelligence at Millward Brown as an Account Coordinator. Years later, she returned to ITESM with the goal of pursuing a postgraduate degree. Today she is the Bachelor Director of Marketing at Tecnológico de Monterrey Campus Mexico City. Since 2017 she has been a professor at the Business School. Students acknowledge her balanced vision of the business world, as well as the care she takes in the professional and personal development of students.

itzelcoquis@tec.mx
https://orcid.org/0000-0001-6284-8320

Mario Iván Contreras Valdez studied for a bachelor's degree in economics and finance at Tecnológico de Monterrey (ITESM) and graduated with honors. Subsequently, he entered EGADE Business School to study for a Ph.D. in Finance, where he obtained recognition for an outstanding thesis. He is currently the Bachelor Director of Finance at Tecnológico de Monterrey Campus Mexico City. Additionally, he is a consultant to firms on topics of M&A and Risk Management. His research areas include cryptocurrencies, quantitative finance, and risk management.

marioivan.contrerasv@tec.mx
https://orcid.org/0000-0003-4870-4376