https://doi.org/10.5154/r.rga.2023.73.14

Food access of poor households in Mexico: a classification tree application

Fernando Elester Vázquez-Ayanegui[1]*
Juan Hernández-Ortiz[1]
Ramón Valdivia-Alcalá[1]
César Botello-Aguillón[2]
Federico Augusto Toledo-Ruiz[1]

Abstract

Poverty and lack of food access are critical problems in developing countries, affecting more than 20 % of the population in Mexico. This research aimed to analyze the factors influencing food access in Mexican households and to develop a predictive model. The machine learning technique, known as classification tree, was used and the results were compared with those of a logit model, frequently used in the literature. In terms of accuracy, the classification tree outperformed the logit model in identifying households at risk of food insecurity (0.6039 *vs.* 0.5402) and provided a visual interpretation of the results. The findings suggest that households living in poverty, located in urban areas, with more than three members or without basic education, should be prioritized in social policies, since they are more likely to face food access problems in Mexico.

Keywords: Social policies, monetary deprivation, welfare lines, food basket, hunger.

El acceso a los alimentos de los hogares pobres en México: una aplicación de árbol de clasificación

Resumen

La pobreza y el acceso a los alimentos son problemas críticos en los países en desarrollo, y afectan a más del 20 % de la población en México. Esta investigación tuvo como objetivo analizar los factores que influyen en el acceso a los alimentos en hogares mexicanos y desarrollar un modelo predictivo. Se empleó la técnica de *machine learning*, conocida como árbol de clasificación, y los resultados se compararon con los de un modelo *logit*, utilizado con frecuencia en la literatura. En términos de precisión, el árbol de clasificación superó al modelo *logit* al identificar hogares en riesgo de inseguridad alimentaria (0.6039 *vs.* 0.5402) y brindó una interpretación visual de los resultados. Los hallazgos sugieren que los hogares en situación de pobreza, ubicados en zonas urbanas, con más de tres integrantes o sin educación básica, deberían ser prioritarios en las políticas sociales, ya que presentan mayores probabilidades de enfrentar problemas de acceso a los alimentos en México.

Palabras clave: Políticas sociales, carencias monetarias, líneas de bienestar, canasta alimentaria, hambre.

[1]Universidad Autónoma Chapingo. Carretera México-Texcoco km 38.5, Chapingo, Estado de México, C. P. 56230, México.
[2]Colegio de Postgraduados. Carretera México-Texcoco km 36.5, Montecillo, Texcoco, Edo. de México, C. P. 56230, México.
*Corresponding author: ayanegui20@gmail.com, tel. 55 320 144 14.

Received: October 27, 2023
Accepted: Novembre 05, 2024

Introduction

Poverty in Mexico represents one of the greatest obstacles to its development. According to data from the National Council for the Evaluation of Social Development Policy (CONEVAL, 2021), more than 40 % of the Mexican population lives in poverty, a situation that has worsened in recent years in part due to the disruptive effects of the SARS-CoV-2 pandemic on the global economy. The proportion of the population living in poverty conditions rose from 41.9 % in 2018 to 43.9 % in 2020 (CONEVAL, 2021), which translates into an increase in monetary and social deprivation for nearly 4 million people. Considering only the monetary dimension (defined according to CONEVAL's welfare lines[1]), in 2020, 52.8 % of the Mexican population did not have sufficient income to cover their basic needs. Furthermore, 17.2 % lacked not only access to essential services such as health or education, but also the resources to purchase the basic food basket. In that same year, 22.5 % of the Mexican population did not have access to nutritious and quality food (CONEVAL, 2021).

Food access is a fundamental pillar of food security, and is related to factors such as economic growth (Manap & Ismail, 2019; Pourreza et al., 2018; Timmer, 2004), health (Monroy-Torres et al., 2021) and sustainability (Pachapur et al., 2020). In the case of Mexico, Urquía-Fernández (2014) points out that food deficiency, its geographic concentration and the associated productive structure have resulted in low productive growth.

Food access (food security) is determined by the size and rurality of localities, economic income (Mundo-Rosas et al., 2021), the average educational level of households, primary production per capita (Díaz-Carreño et al., 2016), age, language, conditions of the head of household (such as sex or marital status) (Magaña-Lemus et al., 2016), and macroeconomic factors such as inflation and employment rate (Díaz-Carreño et al., 2019).

The studies by Mundo-Rosas et al. (2021), Díaz-Carreño et al. (2016) and Magaña-Lemus et al. (2016)

[1] Welfare lines, defined and calculated by CONEVAL, are indicators that establish minimum levels of basic consumption. This organization uses two lines in its calculations: 1) extreme poverty line by income (LPEI: represents the minimum monetary value necessary to cover the basic food basket per person per month), and 2) poverty line by income (LPI: equivalent to the total monetary value of the food basket plus the non-food basket per person per month).

Introducción

La pobreza en México representa uno de los mayores obstáculos para su desarrollo. De acuerdo con datos del Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL, 2021), más del 40 % de la población mexicana vive en pobreza, situación que se ha agravado en los últimos años, en parte, por los efectos disruptivos de la pandemia del SARS-CoV-2 en la economía mundial. La proporción de la población en condiciones de pobreza pasó de 41.9 % en 2018 a 43.9 % en 2020 (CONEVAL, 2021), lo cual se traduce en un aumento en las carencias monetarias y sociales en cerca de 4 millones de personas. Si se considera únicamente la dimensión monetaria (definida según las líneas de bienestar[1] del CONEVAL), en 2020, el 52.8 % de la población mexicana no contaba con ingresos suficientes para cubrir sus necesidades básicas. Además, el 17.2 % carecía no solo de acceso a servicios esenciales como salud o educación, sino que también de recursos para adquirir la canasta básica alimentaria. En ese mismo año, el 22.5 % de la población mexicana no tenía acceso a alimentos nutritivos y de calidad (CONEVAL, 2021).

El acceso a los alimentos es un pilar fundamental de la seguridad alimentaria, y está relacionado con factores como el crecimiento económico (Manap & Ismail, 2019; Pourreza et al., 2018; Timmer, 2004), la salud (Monroy-Torres et al., 2021) y la sustentabilidad (Pachapur et al., 2020). Para el caso de México, Urquía-Fernández (2014) señala que la carencia alimentaria, su concentración geográfica y la estructura productiva asociada se ha traducido en un bajo crecimiento productivo.

El acceso a los alimentos (seguridad alimentaria) está determinado por el tamaño y la condición de ruralidad de las localidades, el ingreso económico (Mundo-Rosas et al., 2021), el nivel educativo promedio de los hogares, la producción primaria per cápita (Díaz-Carreño et al., 2016), la edad, la lengua, las condiciones del jefe de familia (como sexo o situación conyugal) (Magaña-Lemus et al., 2016, y factores macroeconómicos como la inflación y la tasa de empleo (Díaz-Carreño et al., 2019).

[1] Las líneas de bienestar, definidas y calculadas por el CONEVAL, son indicadores que establecen niveles mínimos de consumo básico. Este organismo utiliza dos líneas en sus cálculos: 1) línea de pobreza extrema por ingresos (LPEI: representa el valor monetario mínimo necesario para cubrir la canasta básica alimentaria por persona al mes), y 2) línea de pobreza por ingresos (LPI: equivalente al valor monetario total de la canasta alimentaria más la no alimentaria por persona al mes).

share household-level data and apply multiple categorical response models, such as multivariate logistic or ordered probit. These models require a preselection of covariates or explanatory variables, which can lead to the omission of important variables or the inclusion of some with low explanatory power.

Considering the above, this research aimed to analyze the factors that influence access to food in Mexican households using a machine learning model that allows identifying the variables that provide the best predictive basis for simulating scenarios. To do this, recent and nationally representative information from CONEVAL and the National Household Income and Expenditure Survey (ENIGH) of the National Institute of Statistics, Geography and Informatics (INEGI) was used, in accordance with Magaña-Lemus et al. (2016).

Methodology

Database and treatment

Two cross-sectional databases were used: the CONEVAL 2020 database, hereinafter CONE2020[2] (CONEVAL, 2022), and the ENIGH database for the same year, hereinafter ENIGH2020 (INEGI, 2022). The former provides information on variables related to poverty in Mexican households, and the latter provides a statistical overview of the behavior of household income and expenditures; in addition, it provides information on the occupational, sociodemographic and food access characteristics of household members (INEGI, 2022). The two are compatible and can be integrated through household identifiers.

For the analysis, we used the household-focused section (*concentradohogar*) of ENIGH2020, which groups the most relevant variables, and linked it to the CONE2020 data according to that identifier. Table 1 details the main variables considered in the research.

The dichotomous variable of access to nutritious and quality food is the one used by CONEVAL (2021), and was calculated in two steps. First, the Latin American food security scale was used to identify the level of household insecurity (severe, moderate, mild food insecurity and food security), for which the variables

[2] This is the database used in Mexico's official statistics for poverty issues, and is the result of the institute's processing of the ENIGH for each year.

Los estudios de Mundo-Rosas et al. (2021), Díaz-Carreño et al. (2016) y Magaña-Lemus et al. (2016) comparten datos a nivel de hogar y aplican modelos de respuesta categórica múltiple, como el logístico multivariado o el "*ordered probit*". Estos modelos requieren una preselección de covariables o variables explicativas, lo cual puede dar lugar a la omisión de variables importantes o a la inclusión de algunas con bajo poder explicativo.

Considerando lo anterior, el objetivo de esta investigación fue analizar los factores que influyen en el acceso a los alimentos en hogares mexicanos mediante un modelo de aprendizaje automático (*machine learning*) que permita identificar las variables que proporcionen la mejor base predictiva para la simulación de escenarios. Para ello, se utilizó información reciente y representativa a nacional proveniente del CONEVAL y de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) del Instituto Nacional de Estadística, Geografía e informática (INEGI), en concordancia con Magaña-Lemus et al. (2016).

Metodología

Base de datos y tratamiento

Se emplearon dos bases de datos de corte transversal: la del CONEVAL 2020, en adelante CONE2020[2] (CONEVAL, 2022), y la ENIGH para el mismo año, en adelante ENIGH2020 (INEGI, 2022). La primera provee información sobre variables relacionadas con la pobreza en los hogares mexicanos, y la segunda proporciona un panorama estadístico del comportamiento de los ingresos y gastos de los hogares; además, brinda información sobre las características ocupacionales, sociodemográficas y de acceso a la alimentación de los miembros del hogar (INEGI, 2022). Ambas son compatibles y se pueden integrar mediante identificadores de hogar.

Para el análisis, se utilizó el apartado de concentrado del hogar (concentradohogar) de la ENIGH2020, el cual agrupa las variables más relevantes, y se vinculó con los datos del CONE2020 de acuerdo con dicho identificador. En el Cuadro 1 se detallan las principales variables consideradas en la investigación.

[2] Es la base de datos empleada en las estadísticas oficiales de México para temas de pobreza, y es resultado del procesamiento de la ENIGH por parte del instituto para cada año.

Table 1. Variables used in the research.
Cuadro 1. Variables utilizadas en la investigación.

Variable	Type / Tipo	Description / Descripción
qci / ict	Continuous / Continua	Quarterly current household income / Ingreso corriente trimestral del hogar
qcpci / ictpc	Continuous / Continua	Quarterly current per capita income of the members of a particular household / Ingreso corriente trimestral per cápita de los integrantes de un hogar particular
rururb	Dichotomous / Dicotómica	rururb = 1 if the household is rural, rururb = 0 otherwise / rururb = 1 si el hogar es rural, rururb = 0 en otro caso
houssize / tamhogesc	Continuous / Continua	Household members / Integrantes del hogar
minors_share / menores_share	Continuous / Continua	Share of household members under 16 years of age / Proporción de integrantes del hogar menores de 16 años
ih / hli	Dichotomous / Dicotómica	ih = 1 if the household is considered indigenous, ih = 0 otherwise / hli = 1 si el hogar es considerado como indígena, hli = 0 en otro caso
iph / plp	Dichotomous / Dicotómica	iph = 1 if the household is income poor, iph = 0 otherwise / plp = 1 si el hogar es pobre por ingresos, plp = 0 en otro caso
sex_head / sexo_jefe	Dichotomous / Dicotómica	sex_head = 1 if the head of the family is a woman, sex_head = 0 if it is a man / sexo_jefe = 1 si el jefe de la familia es mujer, sexo_jefe = 0 si es hombre
educa_head / educa_jefe	Discrete / Discreta	No education = 0, basic (some level of primary or secondary education) = 1, high school or professional = 2 / Sin instrucción = 0, básica (con algún nivel de primaria o secundaria) = 1, preparatoria o profesional = 2
acc_food1 / acc_alim1	Dichotomous / Dicotómica	Concern about food running out = 1 / Preocupación porque la comida se acabe = 1
acc_food2 / acc_alim2	Dichotomous / Dicotómica	No food = 1 / Sin comida = 1
acc_food3 / acc_alim3	Dichotomous / Dicotómica	Little variety of food = 1 / Poca variedad de alimentos = 1
acc_food5 / acc_alim5	Dichotomous / Dicotómica	Adult stopped having or eating some of his/her food = 1 / Adulto dejó de tener o ingerir alguno de sus alimentos = 1
acc_food6 / acc_alim6	Dichotomous / Dicotómica	Adult consumes less food = 1 / Adulto consume menos alimentos = 1
acc_food7 / acc_alim7	Dichotomous / Dicotómica	Adult felt hungry and does not consume food = 1 / Adulto sintió hambre y no consume alimentos = 1
acc_food8 / acc_alim8	Dichotomous / Dicotómica	Adult stopped eating at some point = 1 / Adulto dejó de comer alguna vez = 1
acc_food9 / acc_alim9	Dichotomous / Dicotómica	Adult begged for food = 1 / Adulto mendigó por comida = 1
acc_food10 / acc_alim10	Dichotomous / Dicotómica	Minor with unhealthy food = 1 / Menor con alimentos no sanos = 1
acc_food11	Dichotomous / Dicotómica	Minor with little variety of food = 1 / Menor con poca variedad de alimentos = 1
acc_food12	Dichotomous / Dicotómica	Minor consumes less food = 1 / Menor consume menos alimentos = 1
acc_food13	Dichotomous / Dicotómica	Food decreased for a minor = 1 / Disminuyó comida para algún menor = 1
acc_food14	Dichotomous / Dicotómica	Minor felt hungry and does not consume food = 1 / Menor sintió hambre y no consume alimentos = 1
acc_food15	Dichotomous / Dicotómica	Minor went to bed hungry = 1 / Menor se acostó con hambre = 1
acc_food16	Dichotomous / Dicotómica	Minor with one or fewer meals per day = 1 / Menor con una o menos comidas al día = 1

Source: prepared by authors with data from INEGI (2022) and CONEVAL (2022).
Fuente: elaboración propia con datos del INEGI (2022) y CONEVAL (2022).

acc_food1-acc_food16 were added according to the following relationship:

$$security\_level = \begin{cases} \sum_{i=1}^{9} acc_{food_i} if\ the\ household\ has\ no\ minors \\ \sum_{i=1}^{16} acc_{food_i} if the\ household\ has\ minors \end{cases}$$

The variable ranges from 0 to 16, and depends on the number of affirmative responses regarding their food access status. Once the variable was constructed, it was used to classify the degree of food security according to the guidelines of the Latin American and Caribbean Food Security Scale (ELCSA) (Table 2).

The variable *security level* was considered to generate an indicator variable called *without food access*, which was equal to 1 if the households had a severe or moderate degree of food insecurity, and 0 if it was otherwise.

For the implementation of the statistical technique, null values were omitted and the observations were randomly divided into two sets: 70 % were assigned to the training set, used to calibrate the classification tree model, and the remaining 30 % formed the test set, used to evaluate the performance of the model in predicting the response variable.

Statistical model

In the context of machine learning, classifying observations into categories according to a set of attributes (such as determining whether a household has access to food based on its income or character-

La variable dicotómica de acceso a los alimentos nutritivos y de calidad es la empleada por el CONEVAL (2021), y se calculó en dos pasos. Primero, se empleó la escala latinoamericana de seguridad alimentaria para identificar el nivel de inseguridad de los hogares (inseguridad alimentaria severa, moderada, leve y seguridad alimentaria), para lo cual se agregaron las variables acc_alim1-acc_alim16 de acuerdo con la siguiente relación:

$$nivel\_seguridad = \begin{cases} \sum_{i=1}^{9} acc\_alim_i\ si\ el\ hogar\ no\ tiene\ menores \\ \sum_{i=1}^{16} acc\_alim_i\ si\ el\ hogar\ tiene\ menores \end{cases}$$

El rango de la variable va de 0 a 16, y depende del número de respuestas afirmativas sobre su estatus de acceso a los alimentos. Una vez construida la variable, se utilizó para clasificar el grado de seguridad alimentaria de acuerdo con las directrices de la Escala Latinoamericana y Caribeña de Seguridad Alimentaria (ELCSA) (Cuadro 2).

La variable *nivel_seguridad* se consideró para generar una variable indicadora denominada *sin_acceso_alimentos*, la cual fue igual a 1 si los hogares tenían un grado de inseguridad alimentario severo o moderado, y 0 si fue de otra forma.

Para la implementación de la técnica estadística, se omitieron los valores nulos y se dividieron aleatoriamente las observaciones en dos conjuntos: el 70 % se asignó al conjunto de entrenamiento, utilizado para calibrar el modelo de árbol de clasificación,

Table 2. Cut-off points for the classification of food security according to the type of household.
Cuadro 2. Puntos de corte para la clasificación de la seguridad alimentaria según el tipo de hogar.

Type of household / Tipo de hogar	Security / Seguridad	Mild insecurity / Inseguridad leve	Moderate insecurity / Inseguridad moderada	Severe insecurity / Inseguridad severa
Households made up of adults only / Hogares integrados solamente por adultos	0	1 to 3	4 to 6	7 to 9
Households made up of adults and minors / Hogares integrados por adultos y menores	0	1 to 5	6 to 10	11 to 15

Source: Food and Agriculture Organization of the United Nations (FAO, 2012).
Fuente: Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO, 2012).

istics) is part of the supervised techniques. Among the most common classification algorithms are logistic regression, k-Nearest Neighbors and decision trees (Alsharkawi et al., 2021). Their performance depends on the type of information analyzed, the decision criteria and the computational cost.

In this research, a classification tree model was employed to analyze food access for three main reasons: 1) it provides a graphical representation of the results, which is useful for public policy makers by facilitating interpretation, 2) it allows for a pre-selection of the most significant variables to predict the response variable, and 3) in general, it has been shown to have a superior performance to that of logistic regression, which is widely used in similar studies (Alsharkawi et al., 2021; Bagheri & Saadati, 2019; Speybroeck et al., 2004).

A classification tree (CART) is a non-parametric model that generates a binary decision tree. As noted by Speybroeck et al. (2004), the construction of a classification tree starts with a root node containing all available attributes; then, through a process of yes/no questions, descendant nodes are generated. The algorithm identifies the best variable to split the node into two child nodes by evaluating all possible splitters and values of the selected variable. The criterion for choosing the best splitter is to maximize the average "purity" of the child nodes.

The results of a CART model can be visualized through a tree-like structure, in which the *N* observations are classified at each *m* mode and minimizes the following cost function:

$$C_{\propto}(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \propto |T|$$

where $|T|$ is the number of terminal nodes, N_m is the number of observations in the region bounded by node *m*, $Q_m(T)$ is the impurity measure at node *T*, and is the calibration parameter that governs the balance between tree size and its goodness-of-fit to the data. Large values of α generate small trees, while low values form large trees (Hastie et al., 2008).

Among the measures of impurity ($Q_m(T)$), the most notable are the misclassification error, the Gini index and cross-entropy. In this research, the Gini in-

y el 30 % restante conformó el conjunto de prueba (*test*), empleado para evaluar el desempeño del modelo en la predicción de la variable respuesta.

Modelo estadístico

En el contexto del aprendizaje automático de maquina (*machine learning*), la clasificación de las observaciones en categorías de acuerdo con una serie de atributos (como determinar si un hogar tiene acceso a los alimentos en función de sus ingresos o características) forma parte de las técnicas supervisadas. Entre los algoritmos de clasificación más comunes se encuentran la regresión logística, los vecinos más cercanos (*k-Nearest Neighbors*) y los árboles de decisión (Alsharkawi et al., 2021). Su desempeño depende del tipo de información analizada, los criterios de decisión y el costo computacional.

En esta investigación, se empleó un modelo de árbol de clasificación para analizar el acceso a los alimentos por tres razones principales: 1) proporciona una representación gráfica de los resultados, lo cual es útil para los responsables de diseñar políticas públicas al facilitar la interpretación, 2) permite realizar una preselección de las variables más significativas para predecir la variable respuesta, y 3) en general, ha demostrado tener un desempeño superior al de la regresión logística, que es ampliamente usada en estudios similares (Alsharkawi et al., 2021; Bagheri & Saadati, 2019; Speybroeck et al., 2004).

Un árbol de clasificación (CART, por sus siglas en inglés) es un modelo no paramétrico que genera un árbol de decisión binario. Como señalan Speybroeck et al. (2004), la construcción de un árbol de clasificación comienza con un nodo raíz que contiene todos los atributos disponibles; después, mediante un proceso de preguntas de tipo sí/no, se generan nodos descendientes. El algoritmo identifica la mejor variable para dividir el nodo en dos nodos secundarios, mediante la evaluación de todas las posibles divisiones (*splitters*) y valores de la variable seleccionada. El criterio para elegir el mejor divisor es maximizar la "pureza" promedio de los nodos secundarios.

Los resultados de un modelo CART se pueden visualizar mediante una estructura en forma de árbol, en la cual las *N* observaciones se clasifican en cada nodo *m* y minimiza la siguiente función de costos:

dex was used to expand the tree due to its effectiveness in numerical optimization. To guide the reduction of cost complexity, the misclassification error was used, since it has less sensitivity to node assignment (Hastie et al., 2008). Once the classification tree was estimated, its performance was evaluated using two metrics:

$$accuracy = \frac{2{,}303 + 8{,}468}{17{,}837} = 0.6039$$

1. Accuracy measure: identifies the percentage of success in classifying categories, where a true positive (or negative) corresponds to those observations in which the estimate of household access (or lack of access) to quality food coincides with the observed value. In contrast, a false positive (or negative) refers to those observations in which the estimate of access (or lack of access) to quality food differs from its observed value.
2. Confusion matrix: shows the number of correct and incorrect classifications of the test data set in each class.

Results and discussion

For the implementation of the classification tree, the variables listed in Table 1 were used. Although the proportion of observations where the response variable (*without food access*) takes the value of 1 is close to 30 %, it was considered acceptable. Therefore, subsampling or oversampling methods were not applied, which reduce the bias of the results due to the existence of under-represented categories.

Table 3 presents the descriptive statistics of the database used in the research, and provides an overview of the characteristics and distribution of the variables analyzed.

On average, the analyzed households had a quarterly income of $14,827.32 MXN, equivalent to a per capita income of $5,317.00 MXN; however, 50 % of Mexicans have a quarterly income below $3,553.00 MXN. Both values are below the income poverty line established by CONEVAL for 2020, which was $10,857.00 MXN in urban areas and $7,665.00 MXN in rural areas. This indicates that 46.23 % of the population will experience income poverty. In general, 21.56 % of households are clas-

$$C_{\propto}(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \propto |T|$$

donde $|T|$ es el número de nodos terminales, N_m es el número de observaciones en la región delimitada por el nodo m, $Q_m(T)$ es la medida de impureza en el nodo T, y es el parámetro de calibración que rige el equilibrio entre el tamaño del árbol y su bondad de ajuste a los datos. Los valores grandes de α generan árboles pequeños, mientras que valores bajos forman árboles grandes (Hastie et al., 2008).

Entre las medidas de impureza ($Q_m(T)$) destacan el *misclassification error*, el índice de Gini y la *cross-entropy*. En esta investigación, se utilizó el índice de Gini para expandir el árbol, debido a su eficacia en la optimización numérica. Para guiar la reducción de la complejidad de los costos, se utilizó el *misclassification error*, ya que tiene menor sensibilidad a la asignación de nodos (Hastie et al., 2008). Una vez estimado el árbol de clasificación, su desempeño se evaluó mediante dos métricas:

$$accuracy = \frac{2{,}303 + 8{,}468}{17{,}837} = 0.6039$$

Medida de exactitud (*accuracy*): identifica el porcentaje de éxito en la clasificación de las categorías, donde un verdadero positivo (o negativo) corresponde a aquellas observaciones en las que la estimación de acceso (o la falta de acceso) de los hogares a alimentos de calidad coincide con el valor observado. En contraste, un falso positivo (o negativo) se refiere a las observaciones en las que la estimación de acceso (o la falta de acceso) a alimentos de calidad difiere de su valor observado.

Matriz de confusión: muestra el número de clasificaciones correctas e incorrectas del conjunto de datos de prueba en cada clase.

Resultados y discusión

Para la implementación del árbol de clasificación, se utilizaron las variables mencionadas en el Cuadro 1. Aunque la proporción de observaciones donde la variable respuesta (*sin_acceso_alimentos*) toma el valor de 1 es cercana al 30 %, se consideró aceptable. Por ello, no se aplicaron métodos de *subsampling* u *oversampling*, los cuales reducen el sesgo

Table 3. Descriptive statistics of the database used.
Cuadro 3. Estadísticas descriptivas de la base de datos empleada.

Variable	Average / Promedio	Median / Mediana	Standard deviation / Desviación estándar	Observations / Observaciones
qci / ict	14,827.3200	10,533.1200	22,163.4700	89,006
qcpci / ictpc	5,317.4090	3,553.4360	10,620.0600	89,006
houssize / tamhogesc	3.2216	2.9835	1.5395	89,006
ih / hli	0.08340	-	0.27648	89,006
iph / plp	0.4623	NA	0.4986	89,006
rururb	0.21564	-	0.41127	89,006
sex_head / sexo_jefe	0.2987	NA	0.4577	89,006
without food access / sin_acceso_alimentos	0.23263	-	0.42251	89,006

Source: prepared by the authors with data from INEGI (2022) and CONEVAL (2022). Sample weights from the surveys were used.
Fuente: elaboración propia con datos del INEGI (2022) y CONEVAL (2022). Se emplearon los pesos muestrales de las encuestas.

sified as rural, 8.34 % have at least one member who speaks an indigenous language, 29.87 % have a woman as head of household, and 23.26 % have a lack of food access.

Figure 1 shows the classification tree obtained from the training data, optimized to minimize error through cross-validation. Each node of the tree represents a decision, and the branches are split based on the answers. If the answer is negative, the right branch is followed; if it is affirmative, the left branch, until reaching a final node indicating the most likely outcome for each observation. For example, at the initial node, if a household is not in income poverty (*iph* = 0, left branch of the tree), it is highly likely that it has access to food (*without food access* = 0), a scenario that represents 36 % of the households in the training phase.

If the household is in poverty (right branch of the tree) and its quarterly income is less than $3,957.00 MXN (right branch of the second level of the tree), the most likely outcome is lack of food access, a situation that is intensified if the household is located in an urban area (*rururb* = 0). On the other hand, poor households with quarterly incomes above $3,957.00 MXN are more likely to have access to food (52 % of the population, left branch of the second level of the tree).

Based on the classification tree and its results, the following targeting strategy is proposed if we wish to implement a social program that improves access to quality food in Mexican households:

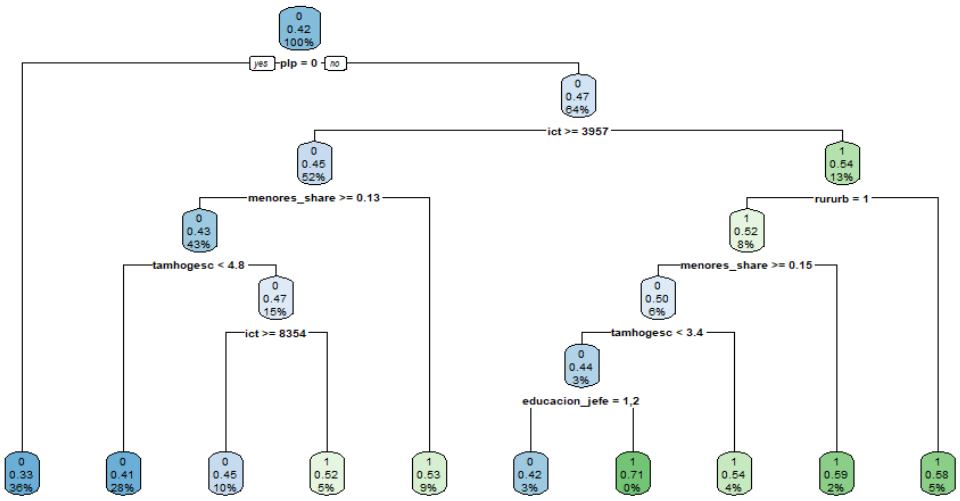
de los resultados ante la existencia de categorías sub-representadas.

El Cuadro 3 presenta las estadísticas descriptivas de la base de datos utilizada en la investigación, y ofrece un panorama general de las características y distribución de las variables analizadas.

En promedio, los hogares analizados tuvieron un ingreso trimestral de $14,827.32 MXN, equivalente a un ingreso per cápita de $5,317.00 MXN; sin embargo, el 50 % de los mexicanos tiene un ingreso inferior a $3,553.00 MXN trimestrales. Ambos valores están por debajo de la línea de pobreza por ingresos establecida por el CONEVAL para 2020, que fue de $10,857.00 MXN en zonas urbanas y de $7,665.00 MXN en zonas rurales. Lo anterior indica que el 46.23 % de la población experimentara pobreza por ingresos. De manera general, el 21.56 % de los hogares se clasifican como rurales, 8.34 % tienen al menos un integrante hablante de lengua indígena, 29.87 % tiene como jefes de familia a una mujer y 23.26 % presenta falta de acceso a los alimentos.

La Figura 1 muestra el árbol de clasificación obtenido a partir de los datos de entrenamiento, optimizado para minimizar el error mediante validación cruzada. Cada nodo del árbol representa una decisión, y las ramas se dividen según las respuestas. Si la respuesta es negativa, se sigue la rama derecha; si es afirmativa, a la rama izquierda, hasta llegar a un nodo final que indica el resultado más probable para cada observación. Por ejemplo, en el nodo inicial, si un hogar no se encuentra en condición de pobreza

Figure 1. Classification tree to determine food access in Mexico. Source: Prepared by the authors.
Figura 1. Árbol de clasificación para conocer el acceso a los alimentos en México. Fuente: elaboración propia.



- Rural households: Select as the first target group rural households with quarterly incomes below $3,957.00 MXN. Within this group, prioritize those with a high share of minors and heads of household with a low educational level.
- Urban households: In urban areas, prioritize households with incomes below $3,957.00 MXN. Pay special attention to households with a significant proportion of minors (more than 13 %) and those whose head of household has a basic level of education or less.
- Household size: Regardless of their location (rural or urban), give priority to households with more than 4.8 members, because household size increases vulnerability to insufficient access to quality food.

Applying these criteria would allow the social program's efforts to be effectively focused on those households most likely to face food insecurity, thereby optimizing the use of available resources and improving the living conditions of the most vulnerable populations.

The absence of some variables from Table 1 in the classification tree is due to the selection of variables with the statistical technique used, which identifies and retains only those variables with the greatest relevance for the prediction. In this case, the variables selected were poverty status, income, education of

por ingresos (*plp* = 0, rama izquierda del árbol), es altamente probable que tenga acceso a alimentos (*sin_acceso_alimentos* = 0), escenario que representa al 36 % de los hogares en la fase de entrenamiento.

Si el hogar se encuentra en situación de pobreza (rama derecha del árbol) y su ingreso trimestral es menor a $3,957.00 MXN (rama derecha del segundo nivel del árbol), el resultado más probable es la falta de acceso a alimentos, situación que se intensifica si el hogar está ubicado en una zona urbana (*rururb* = 0). Por otra parte, los hogares pobres con ingresos trimestrales superiores a $3,957.00 MXN tienen mayores probabilidades de acceder a alimentos (52 % de la población, rama izquierda del segundo nivel del árbol).

Con base en el árbol de clasificación y sus resultados, se propone la siguiente estrategia de focalización si se desea implementar un programa social que mejore el acceso a alimentos de calidad en los hogares mexicanos:

- Hogares rurales: Seleccionar como primer grupo objetivo los hogares rurales con ingresos trimestrales inferiores a $3,957.00 MXN. Dentro de este grupo, priorizar aquellos con una alta proporción de menores y con jefes de hogar con bajo nivel educativo.
- Hogares urbanos: En las zonas urbanas, priorizar a los hogares con ingresos inferiores a $3,957.00 MXN. Prestar especial atención a los

the head of household, and number of members and share of minors in the household. These results are consistent with those reported by Mundo-Rosas et al. (2021) and Díaz-Carreño et al. (2016).

Once the classification tree was estimated, its performance metrics were obtained with the test database. Table 4 shows the confusion matrix, where it can be seen that the success rate in predicting that households do not have access to food reaches 75.05 %; that is, out of every 100 events, 75 of them are correctly predicted.

From the data presented in Table 4, the accuracy indicator (true positives + true negatives among the sample) was calculated. This metric indicates that the estimated model correctly classifies 60.39 % of the cases, covering both households with and without access to food.

In order to compare the performance of the model, a logit model was estimated considering the explanatory variables selected by the classification tree and applying it to the training base. Subsequently, its accuracy metrics were calculated with the test base. The results indicated that, with the exception of the intercept, the estimators are statistically different from zero at 95 % confidence (p-values) (Table 5).

As in the classification tree, poverty status and a higher number of household members increase the probability of not having access to food, while the opposite is true for the rest of the variables.

The interpretation of the logit model coefficients is related to the marginal effects of the explanatory

hogares con una proporción significativa de menores (más del 13 %) y a aquellos cuyo jefe de hogar posea un nivel educativo básico o menor.

- Tamaño del hogar: Independientemente de su ubicación (rural o urbana), dar prioridad a los hogares con más de 4.8 miembros, debido a que el tamaño del hogar incrementa la vulnerabilidad frente al acceso insuficiente a alimentos de calidad.

Aplicar estos criterios permitiría enfocar eficazmente los esfuerzos del programa social en los hogares con mayor probabilidad de enfrentar inseguridad alimentaria, con lo cual se optimizaría el uso de los recursos disponibles y mejoraría las condiciones de vida de las poblaciones más vulnerables.

La ausencia de algunas variables del Cuadro 1 en el árbol de clasificación se debe a la selección de variables con la técnica estadística empleada, la cual identifica y retiene únicamente aquellas variables con mayor relevancia para la predicción. En este caso, las variables seleccionadas fueron la condición de pobreza, el ingreso, la educación del jefe de familia, el número de integrantes y la proporción de menores en el hogar. Estos resultados son consistentes con lo reportado por Mundo-Rosas et al. (2021) y Díaz-Carreño et al. (2016).

Una vez estimado el árbol de clasificación, se obtuvieron sus métricas de desempeño con la base de datos de prueba. En el Cuadro 4 se presenta la matriz de confusión, donde se observa que la tasa de

Table 4. Confusion matrix of the results.
Cuadro 4. Matriz de confusión de los resultados.

		Observed / Observado	
		With access / Con acceso	**Without access / Sin acceso**
Predicted / Predicho	With access / Con acceso	8,468	5,195
	Without access / Sin acceso	1,871	2,303
Error rate / Tasa de error		0.5025	0.2495
Success rate / Tasa de éxito		0.4975	0.7505

Source: prepared by the authors.
Fuente: elaboración propia.

Table 5. Logit model results.
Cuadro 5. Resultados de modelo *logit*.

	Estimator / Estimador	Standard error / Error estándar	z value	Pr(>|z|)
intercept / intercepto	-9.043e-02	7.623e-02	-1.1860	0.2355
iph / plp	2.017e-01	4.007e-02	5.0350	4.77e-07
qci / ict	-3.084e-05	2.509e-06	-12.291	< 2e-16
education_head =1 / *educación_jefe* =1	-1.435e-01	5.317e-02	-2.6990	0.0070
education head =2 / *educación_jefe* =2	-3.100e-01	5.962e-02	-5.2000	1.99e-07
houssize / tamhogesc	8.797e-02	1.084e-02	8.1180	4.73e-16
rururb	-9.386e-02	2.787e-02	-3.3670	0.0008
minors_share / menores_share	-5.980e-01	7.034e-02	-8.5020	< 2e-16
accuracy = 0.5402 / *accuracy* = 0.5402				

Source: prepared by the authors. A value of 0.5 was used as a cut-off point to determine the classification between access and lack of access to food.

Fuente: elaboración propia. Se utilizó un valor de 0.5 como punto de corte para determinar la clasificación entre acceso y falta de acceso a los alimentos.

variables on the logarithm of the probability of not having access to food versus having access. These coefficients, known as log-odds ratios, represent the logarithm of the conditional expectation of observing y = 1 (given the control variables) divided by the logarithm of the conditional expectation of observing y = 0 (given the control variables). Due to the technical complexity associated with interpreting these results versus the visual clarity of the classification tree, the latter is more useful. In addition, the accuracy of the logit model is lower than that of the classification tree, which reinforces the reliability of the latter in identifying the response variable and facilitates its interpretation. For these reasons, the classification tree would be the most desirable tool to support decision making.

Conclusions

Machine learning techniques have become important tools for predicting phenomena or events in various fields and disciplines, including economics. In this context, this research uses a classification tree to identify the main factors influencing access to food by Mexican households. This approach aims to help public policy makers accurately identify the tar-

éxito al predecir que los hogares no tengan acceso a alimentos alcanza el 75.05 %; es decir, de cada 100 eventos se acierta en 75 de ellos.

A partir de los datos presentados en el Cuadro 4, se calculó el indicador de precisión o *accuracy* (verdaderos positivos + verdaderos negativos entre la muestra). Esta métrica indica que el modelo estimado clasifica de manera correcta el 60.39 % de los casos, abarcando tanto los hogares con acceso como aquellos sin acceso a alimentos.

Con la finalidad de comparar el desempeño del modelo, se estimó un modelo *logit* considerando las variables explicativas seleccionadas por el árbol de clasificación y aplicándolo a la base de entrenamiento. Posteriormente, se calcularon sus métricas de precisión con la base de prueba. Los resultados indicaron que, con excepción del intercepto, los estimadores son estadísticamente distintos a cero al 95 % de confianza (p-valores) (Cuadro 5).

Al igual que en el árbol de clasificación, la condición de pobreza y mayor número de integrantes del hogar aumentan la probabilidad de no tener acceso a los alimentos, caso contrario para el resto de las variables.

La interpretación de los coeficientes del modelo *logit* está relacionada con los efectos marginales de

get population when designing strategies related to food security.

The results highlight that poverty status, income level, education of the head of household, number of members and share of minors are the variables with the greatest predictive capacity. Compared to a logit model, the classification tree is easier to interpret and provides greater accuracy in identifying households in a situation of food vulnerability.

In general terms, households most likely to face food insecurity are characterized by being in a condition of income poverty, having a quarterly income of less than $3,957.00 MXN, having more than three members in the household, having low educational levels or living in an urban area.

End of English version

References / Referencias

Alsharkawi, A., Al-Fetyani, M., Dawas, M., Saadeh, H., & Alyaman, M. (2021). Poverty classification using machine learning: The case of Jordan. *Sustainability, 13*(3), 1412. https://doi.org/10.3390/su13031412

Bagheri, A., & Saadati, M. (2019). Modelling childbearing desire: comparison of logistic regression and classification tree approaches. *Crescent Journal of Medical and Biological Sciences*, *6*(4), 487-493.

Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL). (2021). *Nota técnica sobre la medición multidimensional de la pobreza, 2018-2020*. CONRVAL. https://www.coneval.org.mx/Medicion/MP/Documents/MMP_2018_2020/Notas_pobreza_2020/Nota_tecnica_medicion_multidimensional_de_la_pobreza_2018_2020.pdf

Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL). (2022). *Programas de cálculo y bases de datos 2016-2020*. CONEVAL. https://www.coneval.org.mx/Medicion/Paginas/PobrezaInicio.aspx

Díaz-Carreño, M. Á., Díaz-Bustamente, A., & Sánchez-León, M. (2016). Inseguridad alimentaria en los estados de México: un estudio de sus principales determinantes. *Economía, Sociedad y Territorio, 16*(51), 459-483. https://doi.org/10.22136/est002016818

Díaz-Carreño, M. Á., Sánchez-Cándido, L. V., & Herrera Rendón-Nebel, M. T. (2019). La inseguridad alimentaria severa en

las variables explicativas sobre el logaritmo de la probabilidad de no tener acceso a los alimentos versus tener acceso. Estos coeficientes, conocidos como *log-odds ratios*, representan el logaritmo de la esperanza condicional de observar y = 1 (dadas las variables control) entre el logaritmo de la esperanza condicional de observar y = 0 (dadas las variables control). Debido a la complejidad técnica asociada con la interpretación de estos resultados frente a la claridad visual del árbol de clasificación, este último resulta más útil. Además, el *accuracy* del modelo *logit* es inferior al del árbol de clasificación, lo cual refuerza la confiabilidad del segundo para identificar la variable respuesta y facilita su interpretación. Por estas razones, el árbol de clasificación sería la herramienta más deseable para apoyar la toma de decisiones.

Conclusiones

Las técnicas de *machine learning* se han convertido en herramientas importantes para la predicción de fenómenos o eventos en diversos ámbitos y disciplinas, incluida la economía. En este contexto, la presente investigación emplea un árbol de clasificación para identificar los principales factores que influyen en el acceso a los alimentos por los hogares mexicanos. Este enfoque tiene como objetivo facilitar a los hacedores de políticas públicas la identificación precisa de la población objetivo en el diseño de estrategias relacionadas con la seguridad alimentaria.

Los resultados destacan que la condición de pobreza, el nivel de ingreso, la educación del jefe de familia, el número de integrantes y la proporción de menores son las variables con mayor capacidad predictiva. En comparación con un modelo *logit*, el árbol de clasificación es más fácil de interpretar y provee mayor precisión para identificar a los hogares en situación de vulnerabilidad alimentaria.

En términos generales, los hogares con mayor probabilidad de enfrentar inseguridad alimentaria se caracterizan por estar en condición de pobreza por ingresos, tener ingresos trimestrales menores a $3,957.00 MXN, contar con más de tres integrantes en el hogar, presentar bajos niveles educativos o pertenecer al ámbito urbano.

Fin de la versión en español

los estados de México: Un análisis a partir del enfoque de las capacidades 2008-2014. *Estudios sociales. Revista de Alimentación Contemporánea y Desarrollo Regional, 29*(53), 2-24. http://www.redalyc.org/articulo.oa?id=41760730028

Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO). (2012). *Escala Latinoamericana y Caribeña de Seguridad Alimentaria (ELCSA): Manual de uso y aplicaciones*. FAO. https://www.fao.org/3/i3065s/i3065s.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer Series in Statistics. https://doi.org/10.1007/978-0-387-84858-7

Instituto Nacional de Estadística y Geografía (INEGI). (2022). *Microdatos de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH): 2020 Nueva serie*. INEGI. https://www.inegi.org.mx/programas/enigh/nc/2020/default.html

Magaña-Lemus, D., Ishdorj, A., Rosson, C. P., & Lara-Álvarez, J. (2016). Determinants of household food insecurity in Mexico. *Agricultural and Food Economics, 4*, 10. https://doi.org/10.1186/s40100-016-0054-9

Manap, N. M. A., & Ismail, N. W. (2019). Food security and economic growth. *International Journal of Modern Trends in Social Sciences*, *2*(8), 108-118. https://doi.org/10.35631/IJMTSS.280011

Monroy-Torres, R., Castillo-Chávez, Á., Carcaño-Valencia, E., Hernández-Luna, M., Caldera-Ortega, A., Serafín-Muñoz, A., Linares-Segovia, B., Medina-Jiménez, K., Jiménez-Garza, O., Méndez-Pérez, M., & López-Briones, S. (2021). Food security, environmental health, and the economy in Mexico: lessons learned with the COVID-19. *Sustainability*, *13*(13), 7470. https://doi.org/10.3390/su13137470

Mundo-Rosas, V., Unar-Munguía, M., Hernández-F, M., Pérez-Escamilla, R., & Shamah-Levy, T. (2019). La seguridad alimentaria en los hogares en pobreza de México: una mirada desde el acceso, la disponibilidad y el consumo. *Salud Pública de México, 61*(6), 866-875. https://doi.org/10.21149/10579

Pachapur, P. K., Pachapur, V. L., Brar, S. K., Galvez, R., Le Bihan, Y., & Surampalli, R. Y. (2020). Food security and sustainability. In R. Surampalli, T. Zhang, M. K. Goyal, S. Brar & R. Tyagi (eds.), *Sustainability: Fundamentals and Applications*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781119434016.ch17

Pourreza, A., Geravandi, S., & Pakdaman, M. (2018). Food security and economic growth. *Journal of Nutrition and Food Security, 3*(3), 113-115.

Speybroeck, N., Berkvens, D., Mfoukou-Ntsakala, A., Aerts, M., Hens, N., Van Huylenbroeck, G., & Thys, E. (2004). Classification trees versus multinomial models in the analysis of urban farming systems in Central Africa. *Agricultural Systems, 80*(2), 133-149. https://doi.org/10.1016/j.agsy.2003.06.006

Timmer, P. (2004). *Food security and economic growth: An Asian perspective*. Center for Global Development Working. http://dx.doi.org/10.2139/ssrn.1112795

Urquía-Fernández, N. (2014). La seguridad alimentaria en México. *Salud Pública de México, 56*(1), 92-98.