Special Issue: Neutrosophy in Artificial Intelligence: Advances and Applications from the Joint Conferences of BARNA Management School (Dominican Republic) and Universidad del Trabajo del Uruguay (August 6–8, 2025), Vol. 92, 2025

University of New Mexico

Data mining: predictive model of the socioeconomic impact of COVID-19 on the income of cooperative members validated using Neutrosophic TOPSIS and the Hausdorff Distance

Alexander Palacios Zurita[1,2,*], Rosa Giler Zambrano[1], and Dayron Rumbaut Rangel[1]

[1]Universidad Bolivariana del Ecuador, Guayas, Ecuador. wapalaciosz@ube.edu.ec
[2]Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López, Manabí, Ecuador.

Abstract: The COVID-19 pandemic generated a socioeconomic disruption that affected household income, making it crucial to identify modulating factors in vulnerable contexts. This study aimed to develop a predictive model using data mining to determine the socioeconomic, demographic, and well-being factors that influence the income level of 1,456 members of three cooperatives in Manabí, Ecuador, before, during, and after the lockdown. The Interindustry Standard Process for Data Mining methodology was employed. The survey data were segmented into training and test sets, with oversampling applied to the minority classes to correct for imbalances. Several supervised classification algorithms were trained and evaluated in this study. The results showed exceptional predictive performance, with neural networks achieving accuracies close to 90% and areas under the ROC curve greater than 0.95 in the test set. A negative impact of the lockdown was evident, with an increase in the number of members in the lowest income category. The most influential factors were the educational level, employment status, and perception of well-being. These findings offer cooperatives a tool to identify vulnerable profiles and design targeted support strategies, confirming the effectiveness of data mining in modeling economic vulnerability. Furthermore, validation using the Neutrosophic TOPSIS method and Hausdorff Distance corroborated the robustness of the predictive models by incorporating uncertainty and indeterminacy into the assessment of socioeconomic factors, strengthening the reliability of the results in crisis contexts.

Keywords: Data Mining, Predictive Modeling, Credit Unions, Socioeconomic Factors, Machine Learning.

1. Introduction

The COVID-19 pandemic has caused unprecedented socioeconomic disruption globally, significantly impacting the income and financial stability of millions of households [1, 2, 3]. In the Ecuadorian context, credit union members, many of whom belong to vulnerable sectors or the popular and solidarity economy, face particular economic challenges, as demonstrated by studies on their impact at the household level [4]. This situation underscores the crucial need to identify the factors that influenced variations in income levels during the lockdown.

Data mining, which is defined as the process of organizing, analyzing, and extracting valuable information from large datasets [5,6], is emerging as a key methodology. The authors distinguish between descriptive tasks, which classify data based on historical incidents, and predictive tasks, which use historical data to predict future outcomes. Techniques such as classification, regression, and decision trees are fundamental in financial applications for risk prediction or customer loyalty [5]. Other studies highlight the high learning capacity, intelligence, and precision of data mining

technology, considering it essential for processing complex information and transforming it into tools applicable in various fields [6].

However, despite the potential of advanced analytical tools, the central problem in addressing the income situation of COAC members during the COVID-19 crisis lies in the need for a methodological approach that allows to accurately identify the specific factors that modulated these income levels and facilitates the development of robust predictive models for this scenario. The analysis of economic dynamics in the context of health crises constitutes an essential basis for promoting social stability and economic recovery [3]. Despite the advances in machine learning in economics and finance, and its advantage over traditional statistical models [7], a gap persists in the application of data mining methodologies aimed at studying socioeconomic, demographic, and well-being factors at the individual level during health emergencies. While there are data mining applications in the Ecuadorian cooperative sector, such as the work of Mejía Vanegas [8] to predict credit delinquency, and studies that use machine learning to analyze the socioeconomic impact of COVID-19 at the regional or macro level [7, 1, 9, 3], most have focused on aggregate variables or unsupervised approaches, leaving a gap in the construction of supervised predictive models for individual admission in this specific context.

Traditional approaches often fail to capture complex patterns or build the robust predictive models required. This limits credit unions' ability to anticipate situations of economic vulnerability among their members and design effective support strategies. As Winston points out, socioeconomic status is a determining factor in research on the impact of pandemics [10]. Along these lines, the overall objective of this work is to develop a predictive model based on data mining techniques to identify socioeconomic, demographic, and well-being perception factors associated with the income level of members of three segment 1 credit unions in Manabí province, based on a sample of 1,456 respondents, during the COVID-19 lockdown period. This highlights the importance of this approach for informed decision-making in the cooperative sector. Furthermore, the results of the predictive model are validated using the Neutrosophic TOPSIS method and the Hausdorff Distance , which incorporates neutrosophic logic to assess the robustness of predictions and the relevance of identified factors, considering the uncertainty and indeterminacy inherent in socioeconomic data in crisis contexts.

This article is organized as follows: it begins with a presentation of the pandemic context and the problems that justify the research. It then reviews the relevant literature on economic analysis in crises and the application of data mining. The methodology employed is then detailed, including a description of the data collected from the three cooperatives, the preprocessing , the machine learning algorithms used by the Orange Data Mining software , and the evaluation metrics. The results obtained from the comparative analysis of the predictive models and the identification of the most influential factors on income levels are then presented and discussed. The validation of the results using the Neutrosophic TOPSIS method and Hausdorff Distance is then described to confirm the consistency of the predictions and determining factors under a neutrosophic framework. Finally, the conclusions of the study are presented, along with the limitations and prospects for future research in this area.

2. Materials and methods

This research was developed under a quantitative approach, based on the collection and analysis of structured data to identify patterns, relationships and build predictive models. Following the scientific research process [11], once the problem was posed and the literature reviewed, specific hypotheses were derived and tested using a non-experimental, ex post facto design. This design is justified given that data collected on already occurred events (the socioeconomic effects of the COVID-19 pandemic on income) were analyzed, and for each of the three periods analyzed (before, during and after confinement), the study had a cross-sectional component [12]. Furthermore, the analytical-

synthetic method was used to decompose the income level phenomenon into its constituent variables and then reintegrate them into predictive models, as well as an inductive-deductive approach to infer general patterns from the sample data and test the hypotheses [11].

The scope of the research was initially descriptive, according to Barbosa's definition [13], characterizing the current nature of the sample and the variables in each period, recording and analyzing the present phenomena. Subsequently, it progressed to a correlational-associative level, exploring the relationships between socioeconomic, demographic and well-being perception factors with the income level of the partners. The fundamental component of the study was predictive, focusing on the development and evaluation of machine learning models to estimate the income category of the partners at each of the three time points. Finally, by interpreting the best-performing models, we sought to reach a limited explanatory level, identifying the most determining factors of the income level in each context [12].

2.1 Data analysis

For the processing and analysis of the collected data, the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was adopted, a robust and widely recognized framework for data mining projects. This iterative approach consists of six interrelated phases that guided the development of this study, from the initial understanding of the data to the evaluation of the constructed predictive models [14, 15]. The activities carried out in each of these phases are detailed below:

Understanding the business/problem

The main objective of this phase was to clearly define the problem and research objectives from a practical perspective for credit unions. The goal was to understand the need to identify the factors that modulated members' income levels during the socioeconomic disruption caused by the COVID-19 pandemic, so that these findings could inform strategies to support and mitigate economic vulnerability.

Understanding the data

This phase involved familiarizing themselves with the dataset. The study population consisted of members of savings and credit cooperatives in segment 1, regulated by the Superintendency of Popular and Solidarity Economy, located in the province of Manabí, Ecuador. Of a total of four cooperatives belonging to this segment in said province, three agreed to participate in the study: Cooperativa de Ahorro y Crédito Calceta Ltda , Cooperativa de Ahorro y Crédito Chone Ltda , and Cooperativa de Ahorro y Crédito La Benefica . To preserve the confidentiality of the organizations, the participating cooperatives were anonymized , assigning them the names Coop 1, Coop 2, and Coop 3, in no particular order or relation to the alphabetical order in which they were previously mentioned.

The cooperatives were selected using convenience sampling, considering only those that agreed to participate in the research process. Participants within each cooperative were selected using simple random sampling. A consolidated sample of 1,456 members was used, distributed as follows: 382 members in Cooperative 1, 381 members in Cooperative 2, and 693 members in Cooperative 3.

The main data collection technique was the survey, a procedure commonly used in quantitative research to obtain the necessary information to answer the research question [16]. The specific instrument used was a structured questionnaire with closed questions, designed to capture information on demographic variables (age, gender), socioeconomic variables (educational level, employment status, type of housing, family responsibilities), perception of well-being and average monthly income level [17]. This instrument was applied to collect data corresponding to the periods before, during and after the COVID-19 lockdown, generating a database with a temporal dimension.

The variables that served as input for training the income level predictive models are detailed in Table 1.

Table 1 Variables to model the income level of **partners**

Variable	Description	Variable type	Categories
Log in before Admission during Entry after	Member's average monthly income category before, during, or after the COVID-19 lockdown	Categorical ordinal	Less than 1 SBU Mayor equal to 1 SBU per minor and 2 SBU Mayor equal to 2 SBU per minor and 3 SBU More than 3 SBUs for every 4 SBUs
Level of education	Highest level of education achieved by the member	Categorical ordinal	None Essential High school Technique University Mastery Doctorate
Well-being before Wellbeing during Well-being after	Members' perception of how they lived with their income before, during, and after the lockdown	Categorical ordinal	Living badly Living more or less well Live well
Working before Working during Working after	Indicator of whether the member was working before, during, and after confinement	Categorical dichotomous	BUT
Age	Age range of the couple	Categorical ordinal	Ages 18 to 45 years Ages 46 to 65 years people over 65 years old

Gender	Gender with which the couple identifies	Categorical nominal	Man Women Other
Family responsibilities before Family responsibilities during Family responsibilities after	Number of dependents of the member	Categorical ordinal	0 1 or 2 More than 2
Previously rented homes Housing rented during Rental housing after Own house before Own home during Own house later	Type of housing in which the member resided before, during, and after confinement.	Categorical dichotomous	BUT
Identification Cooperative	Cooperative identifier	Categorical nominal	Coop1 Coop2 Coop3

Data preparation

This phase encompassed all the activities required to construct the final dataset from the raw data. It was performed using Orange Data Mining software (version 3.39) and was divided into the following strands:

- Consolidation and Cleanup: Merge data from the three cooperatives and identify and address inconsistencies.
- Handling missing values: Mode imputation was applied for categorical variables.
- Cooperativa_ID was included to identify the origin of each member and categories of the income variable were merged for each period (resulting in four final categories: Less than SBU; 1-2 SBU; 2-3 SBU; and Greater than or equal to 3 SBU) to ensure an adequate number of instances per class.
- Data format: The data were structured for each of the three temporal analyses (before, during, after), selecting in each case the income variable for the corresponding period as the target variable and the demographic, socioeconomic and well-being characteristics of that same period as predictor variables.

Splitting Training and Testing Data

To ensure an objective assessment and avoid overfitting, the dataset of 1,456 partners was segmented for each temporal analysis (before, during, and after). The widget A data sampler was used to perform a stratified split, assigning 80% of the data to the training set (1,165 instances) and the

remaining 20% to the test set (291 instances). This partitioning was kept constant and replicable to ensure comparability of the results.

Training set balance

Due to the observed imbalance in the distribution of the input variable categories, especially in the "during" and "after" periods, a random oversampling technique was applied to only 80% of the training set. Using the Select Rows, Data Sampler, and Concatenate widgets, instances of minority classes were replicated to a more balanced distribution. This step is essential so that the learning algorithms are not biased toward the majority class and can learn to identify patterns across all income categories. Twenty percent of the test set was kept in its original, unbalanced state to simulate real-world conditions during the evaluation.

Modeling

With the training data prepared and balanced, various supervised classification modeling techniques were selected and applied to build the predictive models. The algorithms evaluated included decision trees, random forests, naive Bayes, logistic regression, support vector machines (SVMs), and neural networks. Separate sets of models were built for each of the three time periods, using the balanced training data from each period. Algorithm parameters were kept at their default settings in Orange for initial comparison. The construction of the predictive model can be seen in Appendix A.

Assessment

The final evaluation of each model's performance for each period was performed on the 20% test set, ensuring an objective assessment of the models' generalization. In the "Test and Score" widget, the "Test on test data" option was selected to perform this evaluation. The performance metrics used included Area Under the ROC Curve (AUC), Overall Accuracy (CA), F1-score, Precision, Recall, and Matthews Correlation Coefficient (MCC) [18]. The statistical comparison of performance between models was based on the evidence provided by the widget. Additionally, confusion matrices of the best-performing models were analyzed to understand their success and error patterns by class. Interpretation of influencing factors was performed using nomograms for Logistic Regression and feature importance analysis with the Rank widget, tools that facilitate understanding of model determinants [15,18].

Deployment (adapted to the research context).

Although this research does not involve the implementation of the models in a productive system, the results and knowledge generated are documented in this article for dissemination to the scientific community and the cooperatives themselves. The findings seek to provide information that COACs can use to better understand their members' vulnerability and to design future support strategies or early warning systems.

Hypothesis formulation

Tables 2 and 3 propose the hypotheses to be tested by evaluating the performance of the models.

Table 2 Null and alternative hypothesis on the best predictive model for each **period**

$H_{0\text{-}1}$	There are no statistically significant differences in the predictive ability of machine learning algorithms to estimate income levels before, during, and after the COVID-19 lockdown.

$H_{1.1}$	At least one machine learning algorithm has significantly greater predictive capacity to estimate income levels in the periods studied.

Table 3. Null and alternative hypotheses on the impact of confinement on predictability and determining factors

$H_{0.2}$	The predictive performance of revenue models and the relevance of socioeconomic factors do not vary significantly between the periods before, during, and after the COVID-19 lockdown.
$H_{1.2}$	The predictive performance of revenue models and the relevance of socioeconomic factors vary significantly across the periods studied, reflecting the impact of the pandemic.

2.2. Validation method.

Neutrosophic, a new branch of philosophy, emerged as a generalization of dialectics and Chinese YinYang philosophy. It delves not only into the dynamics of opposites, but also into the dynamics of opposites in conjunction with their neuters (<A>, <neutA> , <antiA>) , where <A> represents an element, <antiA> its opposite, and <neutA> its neutral state (indeterminacy between them). Neutrosophic emphasizes the importance of neutrality/indeterminacy (<neutA>) , giving rise to concepts such as the neutrosophic set, logic, probability, statistics, and measurement, with various practical applications in various fields. In particular, the components of the single-valued neutrosophic set/logic could sum up to 3, highlighting the independence between these components [22].

1 **Definition 1, [** 23 **], [24]** . Let X a space of basic elements denoted by . $\in X$. A NS A in , is described by a truth membership function $T_A(x)$, an indeterminate membership function $I_A(x)$ and a falsehood membership function $F_A(x)$. These membership functions are standard or non-standard real subsets of $]0^-,1^+[$, i.e., $T_A(x): X \to]0^-,1^+[$, $I_A(x): X \to]0^-,1^+[$ y $F_A(x): X \to]0^-,1^+[$, where $0^-=0-\varepsilon$ and $1^+=1+\varepsilon$, while ε is a number greater than 0. Since NS have a restriction mentioned on the sum of the three membership functions, then $0^- \leq \sup T_A() + \sup I_A() + \sup F_A() \leq 3^+$.

2 To address the challenges associated with the practical application of neutrosophic sets (NEs) in a technical and scientific manner, Wang et al. introduced the concept of single-valued neutrosophic sets (SVUs) [25].

3 **Definition 2. [23],[24]** . Let X be a space of basic elements denoted by $x \in X$. B is a SVNS in X with membership functions: truth $T_B(x)$, indeterminacy $I_B(x)$ and falsity $F_B(x)$, which belong to the interval [0,1]. So for any element x, $T_B(x) \in [0,1]$, $I_B(x) \in [0,1]$ and $F_B(x) \in [0,1]$, therefore it is satisfied that $0 \leq T_B(x) + I_B(x) + F_B(x) \leq 3$. For a SVNS B in X , a triplet b will be represented as $\langle T_B(x), I_B(x), F_B(x)\rangle$ for $x \in X$, or as (T_b, I_b, F_b), in a simplified form, and is defined as a single-valued neutrosophic number (SVNN), an element of the SVNS B in X .

4 **Definition 3. [25].** Let A and B be two SVNS in the space X. It can be stated that B contains A ($A \subseteq B$), if and only if $T_A(x) \leq T_B(x), I_A(x) \geq I_B(x)$ *and* $F_A(x) \geq F_B(x)$for all $x \in X$.

5 **Definition 4. [26].** Let $\{B_1, B_2, \dots, B_n\}$ SVNNs ∈ SVNS(x), the Single Valued Weighted Average Neutrosophic Operator (SVNWAO), is then defined as:

6 $\text{SVNWAO}_w(B_1, B_2, \ldots, B_n) = \langle 1 - \prod_{j=1}^{n}\left(1 - T_{B_j}(x)\right)^{w_j}, \prod_{j=1}^{n}\left(I_{B_j}(x)\right)^{w_j}, \prod_{j=1}^{n}\left(F_{B_j}(x)\right)^{w_j}\rangle$

(1)

7 where $B_j = (T_j, I_j, F_j)$ (j = 1, 2, …, n) and $w = (w_1, w_2, \ldots, w_n)$ is a vector, such $\sum w_j = 1$, and $w_n \in [0,1]$

8 **Definition 5. [27]. The** normalized Hamming distance between a = (T_a, I_a, F_a) and b = (T_b, I_b, F_b), SVNN of the SVNS C at X, is defined as:

9 $$d_{Xu}(a,b) = \frac{1}{3}\{|T_a - T_b| + |I_a - I_b| + |F_a - F_b|\} \quad \textbf{(2)}$$

10 **Definition 6 [28]. The** Hausdorff distance between a = (T_a, I_a, F_a) and b = (T_b, I_b, F_b), SVNN of SVNS C in X, is defined as:

11 $$d_{Xu}(a,b) = max\{|T_a - T_b|, |I_a - I_b|, |F_a - F_b|\} \quad (3)$$

12 The Technique of Preference Ordering by Similarity to Ideal Solution (TOPSIS) is an efficient methodology for addressing complex multi-criteria decision problems, such as quality item prioritization. Based on the principle of identifying ideal and anti-ideal solutions, TOPSIS facilitates a nuanced understanding of the decision space, distinguishing optimal from suboptimal options [29]. Using a similarity metric, it quantitatively measures the desirability of each option, improving objectivity. The integration of neutrosophic elements enhances TOPSIS, adding sophistication to the handling of indeterminate information [30], aligning with the changing landscape of decision science and providing a more robust framework for addressing uncertainties and improving decision-making effectiveness.

13 The ideal SVNNS criteria are calculated as indicated in [30]. Where *B is* the positive type criterion and *C* is the cost type criterion, the ideal SVNNS for each case is calculated by:

14 $$B* = \left(T_{\rho^+w}(\beta_j), I_{\rho^+w}(\beta_j), F_{\rho^+w}(\beta_j)\right) \quad (4)$$

15 Denotes the positive ideal solution, corresponding to *B* .

16 $$C* = \left(T_{\rho^-w}(\beta_j), I_{\rho^-w}(\beta_j), F_{\rho^-w}(\beta_j)\right) \quad (5)$$

17 Denotes the negative ideal solution, corresponding to C.

18 Where

19 $$T_{\rho^+w}(\beta_j) = \begin{cases} max_i\left[T_{\rho_i w}(\beta_j)\right], & if \quad j \in B \\ min_i\left[T_{\rho_i w}(\beta_j)\right], & if \quad j \in C \end{cases} \quad (6)$$

20 $$I_{\rho^+w}(\beta_j) = \begin{cases} min_i\left[I_{\rho_i w}(\beta_j)\right], & if \quad j \in B \\ max_i\left[I_{\rho_i w}(\beta_j)\right], & if \quad j \in C \end{cases} \quad (7)$$

21 $$F_{\rho^+w}(\beta_j) = \begin{cases} min_i\left[F_{\rho_i w}(\beta_j)\right], & if \quad j \in B \\ max_i\left[F_{\rho_i w}(\beta_j)\right], & if \quad j \in C \end{cases} \quad (8)$$

22 AND

23 $$T_{\rho^-w}(\beta_j) = \begin{cases} min_i\left[T_{\rho_i w}(\beta_j)\right], & if \quad j \in B \\ max_i\left[T_{\rho_i w}(\beta_j)\right], & if \quad j \in C \end{cases} \quad (9)$$

24 $$I_{\rho^{-}w}(\beta_j) = \begin{cases} max_i[I_{\rho_i w}(\beta_j)], & if \quad j \in B \\ min_i[I_{\rho_i w}(\beta_j)], & if \quad j \in C \end{cases} \quad (10)$$

25 $$F_{\rho^{-}w}(\beta_j) = \begin{cases} max_i[F_{\rho_i w}(\beta_j)], & if \quad j \in B \\ min_i[F_{\rho_i w}(\beta_j)], & if \quad j \in C \end{cases} \quad (11)$$

26 The distance between each SVNN and both ideal values will be calculated by (3), while the proximity coefficient (PC) will be calculated by:

27 $$PC_j = \frac{D^-}{D^+ + D^-}$$

28 Where

29 $0 \leq PC_j \leq 1$

30 $$D^- = d_{Xu}(\beta_j, C*) \quad (12)$$

31 $$D^+ = d_{Xu}(\beta_j, B*) \quad (1\,3)$$

3. Results

The key sociodemographic characteristics of the 1,456 members surveyed, distributed across three credit unions in the province of Manabí, are described below. The distribution by age, gender, and highest level of education attained will be analyzed to contextualize the study population before examining their income levels.

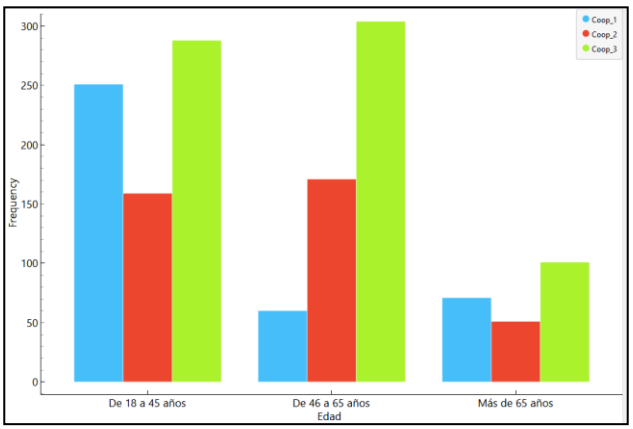

Figure 1 Distribution of partners in cooperatives by age range.

Figure 1 shows a varied age distribution among the three cooperatives. Cooperative 1 (blue) has a higher concentration of members in the 18-45 age range. Cooperative 2 (red) has a more balanced representation in the 18-45 and 46-65 age ranges. Cooperative 3 (green) shows a predominance of members in the 18-45 and 46-65 age ranges, with a lower proportion of members over 65 compared to the other two, although it remains the largest group in this cooperative in the "Over 65" range.

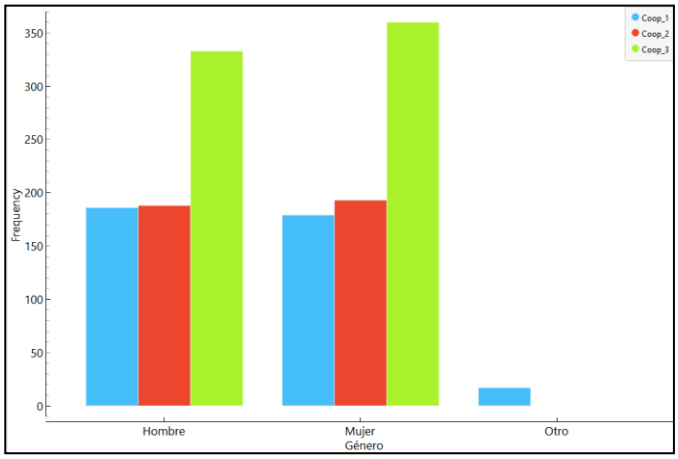

Figure 2 Distribution of partners in cooperatives by gender

As can be seen in Figure 2, the distribution between men and women appears to be relatively balanced in all three cooperatives, although with slight differences. Cooperative 3 (green) shows a higher number of women and men overall due to its larger sample size. The "Other" category has minimal or no representation.

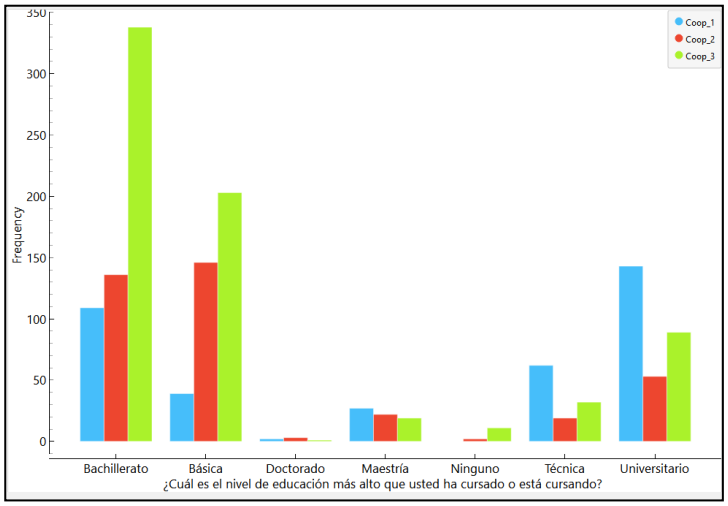

Figure 3education level.

The majority of respondents from the three cooperatives have primary and secondary education, with Coop. 3 concentrating the largest number of people at these levels. Coop. 1 stands out for having a higher proportion of people with technical and university education, suggesting a slightly higher educational profile. Coop. 2 presents a more balanced distribution, although with less representation at the higher levels. Master's and doctoral degrees are practically nonexistent in all three cooperatives, and only a small group has not completed any educational level (Figure 3).

3.1 Analysis of the distribution of the dependent variable (income level) in the three periods

To explain the economic context of the members and the impact of the pandemic, Table 4 presents the distribution of the average monthly income reported during the three study periods: before, during, and after the COVID-19 lockdown. This description will allow us to visualize the changes in the income structure of the total sample.

Table 4 Distribution of members according to income level in the periods studied

		Income before		Income during		Income after	
	COOPERATIVE	**No.**	**In general**	**No.**	**In general**	**No.**	**In general**
Mayor equal to 1 SBU $425.00 monthly and 2 SBU $850.00	Cooperative 1	152	10.44%	132	9.07%	145	9.96%
	Cooperative 2	224	15.38%	187	12.84%	206	14.15%
	Cooperative 3	134	9.20%	131	9.00%	141	9.68%
Mayor equal to 2 SBUs $850.00 monthly and 3 SBUs $1275.00	Cooperative 1	54	3.71%	39	2.68 %	41	2.82 %
	Cooperative 2	18	1.24%	18	1.24 %	23	1.58 %
	Cooperative 3	4	0.27%	4	0.27 %	5	0.34 %
Mayor equal to 3 SBUs $1275.00 monthly and 4 SBUs $1700.00	Cooperative 1	22	1.51%	25	1.72 %	25	1.72 %
	Cooperative 2	2	0.14%	2	0.14 %	3	0.21 %
	Cooperative 3	1	0.07%	1	0.07 %	1	0.07 %
Less than a Unified Basic Wage (UBW)	Cooperative 1	154	10.58%	186	12.77 %	171	11.74 %
	Cooperative 2	137	9.41%	174	11.95 %	149	10.23 %
	Cooperative 3	554	38.05%	557	38.26 %	546	37.50 %
Σ		1456	100%	1456	100%	1456	100%

Before the lockdown, a distribution was observed where the category "Less than one Unified Basic Salary (UBS)" grouped 845 partners (58.04%), while 492 partners (33.79%) were in the range "Greater than or equal to 1 UBS and less than 2 UBS". The highest income categories were considerably lower: 76 partners (5.22%) in "Greater than or equal to 2 UBS and less than 3 UBS" and only 25 partners (1.72%)

in "Greater than or equal to 3 UBS". This indicates that, although a significant portion were already in the lower income levels, there was a considerable segment with lower-middle incomes (Table 4).

During the lockdown, a dramatic increase was observed in the proportion of members in the "Less than SBU" category, which rose to 917 members, representing an alarming 62.98% of the total. This change was mainly due to the "Greater than or equal to 1 SBU and less than 2 SBU" category, which decreased to 450 members (30.91%), and also to the higher categories: "Greater than or equal to 2 SBU and less than 3 SBU" decreased to 61 members (4.19%), and "Greater than or equal to 3 SBU" remained marginally unchanged at 28 members (1.92%). These data reflect a serious negative impact of the pandemic, with widespread impoverishment and a concentration of members in the lowest income stratum (Table 4).

Following the lockdown, the "Less than SBU" category remained the most prevalent, with 866 members (59.48%), a figure still much higher than the pre-pandemic figure, although slightly lower than at the peak of the lockdown. A slight recovery was observed in the "Greater than or equal to 1 SBU and less than 2 SBU" category, which increased to 492 members (33.79%), returning to a level similar to that observed before the pandemic. However, the higher-income categories ("Greater than or equal to 2 SBU and less than 3 SBU", with 69 members or 4.74%, and "Greater than or equal to 3 SBU", with 29 members or 1.99%) showed very limited or no recovery compared to the "During" period, and remained below or barely at the level of their already low pre-pandemic percentages. This suggests that while some members managed to slightly improve their situation after the most critical phase, the negative economic impact persisted for a large majority and the income structure failed to return to pre-crisis levels, with a larger proportion of members consolidating at the lowest income level (Table 4).

3.1. Analysis of the determinants of income level

Prior to evaluating the predictive models, an analysis was conducted to identify the socioeconomic, demographic, and well-being characteristics with the greatest potential influence on determining the members' income level. To do this, Orange's "Rank" widget was used, applying metrics such as Chi-square (χ^2), Information Gain and Gain Ratio, connecting it to the preprocessed data and defining the input variable for each period (before, during, and after lockdown) as the target. This univariate analysis provides an overview of the individual relevance of each characteristic (Figs. 4, 5, and 6).

		#	Inf...ain	Gai...tio	Gini	χ^2
1	¿Cuál es el nivel de educación más alto que usted ha cursado o está cursando?	7	0.201	0.096	0.068	599.018
2	Antes del confinamiento ¿usted se encontraba laborando?	2	0.192	0.203	0.092	122.570
3	Antes del confinamiento por Covid-19, con los ingresos que recibe usted	3	0.175	0.157	0.058	145.955
4	ID_Coop	3	0.175	0.115	0.075	143.449
5	Género	3	0.024	0.022	0.009	9.093
6	Antes del confinamiento, indique el número de cargas familiares	3	0.019	0.016	0.006	6.979
7	Edad	3	0.014	0.010	0.006	20.288
8	Antes del confinamiento ¿Usted tenía una vivienda propia para residir?	2	0.005	0.008	0.000	1.109
9	Antes del confinamiento ¿Usted arrendaba alguna vivienda para residir?	2	0.005	0.007	0.000	4.623

Figure 4 Variables with the greatest influence in the Previous Period

		#	Inf...ain	Gai...tio	Gini	χ^2
1	¿Cuál es el nivel de educación más alto que usted ha cursado o está cursando?	7	0.199	0.096	0.071	609.150
2	Durante el confinamiento ¿usted se encontraba laborando?	2	0.198	0.201	0.092	148.188
3	Durante el confinamiento por Covid-19, con los ingresos que recibe usted	3	0.149	0.130	0.051	62.945
4	ID_Coop	3	0.122	0.080	0.046	102.444
5	Género	3	0.021	0.019	0.009	9.939
6	Durante el confinamiento, indique el número de cargas familiares	3	0.020	0.016	0.006	6.293
7	Edad	3	0.011	0.008	0.006	16.165
8	Durante el confinamiento ¿Usted tenía una vivienda propia para residir?	2	0.001	0.002	0.000	0.396
9	Durante el confinamiento ¿Usted arrendaba alguna vivienda para residir?	2	0.001	0.001	0.000	1.332

Figure 5 Variables with the greatest influence in the Period **During**

		#	Inf...ain	Gai...tio	Gini	χ^2
1	¿Cuál es el nivel de educación más alto que usted ha cursado o está cursando?	7	0.211	0.101	0.074	614.829
2	Después del confinamiento ¿usted se encontraba laborando?	2	0.153	0.162	0.076	99.809
3	ID_Coop	3	0.137	0.090	0.057	110.937
4	Después del confinamiento por Covid-19, con los ingresos que recibe usted	3	0.128	0.115	0.042	83.343
5	Género	3	0.022	0.020	0.009	11.259
6	Después del confinamiento indique el número de cargas familiares	3	0.019	0.015	0.005	5.938
7	Edad	3	0.016	0.011	0.009	24.282
8	Después del confinamiento ¿Usted tenía una vivienda propia para residir?	2	0.002	0.003	0.000	0.596
9	Después del confinamiento ¿Usted arrendaba alguna vivienda para residir?	2	0.001	0.002	0.000	1.947

Figure 6 Variables with the greatest influence in the **Post Period**

Across all three periods analyzed, the "Rank" widget consistently pointed to "**Educational Level**" as the most relevant characteristic for distinguishing income levels, presenting the highest values in metrics such as Chi-square (χ^2) (Before: 599, During: 609, After: 614) and Information Gain (Before: 0.201, During: 0.199, After: 0.211). Immediately after, and also with high consistency, the "**Employment Status**" **variable** for the respective period (e.g., "Were you working before the lockdown?") ranked second, closely followed by **perceptions of well-being and the ID Coop** variable. (cooperative identifier). Variables such as "Age," "Gender," "Number of dependents," and those related to housing tenure or renting showed consistently lower individual influence according to these univariate metrics , although their importance can be reflected in multivariate models. This preliminary hierarchy suggests that educational level, employment status, perception of well-being, and the cooperative to which members belong are, a priori, the factors with the greatest individual discriminatory power for members' income levels in the different temporal contexts studied.

3.2. Predictive models of income level reference

To identify the algorithm with the best ability to predict members' income levels, a set of six supervised classification models was trained and evaluated for each of the three periods. Following the described methodology, the models were trained using 80% of the data (training set), to which an oversampling technique was applied to balance the input classes. Performance evaluation was performed using the remaining 20% of the data (test set), which was not used during training or compensation. Tables 5, 6, and 7 below present the performance results obtained for the periods before, during, and after the lockdown, respectively.

Before the period

Table 5. Predictive performance of 5in the period BEFORE the pandemic

Model	AUC	California	F1	Price	Remember	MCC
Random Forest BEFORE	0.951	0.859	0.855	0.858	0.859	0.732
Naive Bayes BEFORE	0.877	0.756	0.754	0.762	0.756	0.553
Logistic regression BEFORE	0.904	0.825	0.818	0.823	0.825	0.664
SVM BEFORE	0.834	0.643	0.647	0.718	0.643	0.399
Neural network BEFORE	0.961	0.876	0.873	0.873	0.876	0.765
Tree BEFORE	0.944	0.842	0.836	0.837	0.842	0.698

The evaluation of the models for the BEFORE period (Table 5) reveals exceptional predictive performance after applying class balancing and a split between training and testing. The Neural Network model is positioned as the most effective, leading in all key metrics with an AUC of 0.961, an Overall Accuracy (CA) of 0.876, and an MCC of 0.765. It is closely followed by the Random Forest (AUC=0.951, CA=0.859) and the Decision Tree (AUC=0.944, CA=0.842), both with very robust results. These findings demonstrate that it is possible to estimate the pre-pandemic income level with high accuracy using the developed models.

Period During

Table 6. Predictive performance of 6models in the DURING period

Model	AUC	California	F1	Price	Remember	MCC
Tree DURING	0.953	0.863	0.860	0.860	0.863	0.723
Logistic regression DURING	0.919	0.814	0.814	0.824	0.814	0.636
Neural network DURING	0.973	0.904	0.904	0.905	0.904	0.808
SVM DURING	0.878	0.742	0.743	0.756	0.742	0.490
Bayes DURING	0.900	0.777	0.778	0.785	0.777	0.564
Random Forest DURING	0.968	0.890	0.889	0.890	0.890	0.779

During the most acute phase of the lockdown, the predictive models maintained exceptionally high performance, as shown in Table 6. The Neural Network model emerged as the top performer, achieving near-perfect discrimination with an AUC of 0.973 and the highest Overall Accuracy (CA) of 0.904. It was closely followed by the Random Forest (AUC=0.968, CA=0.890) and Decision Tree (AUC=0.953, CA=0.863), whose results were also outstanding. These findings indicate that, despite the economic disruption, socioeconomic factors of the period remained strong predictors of income level, especially after balancing the training data.

Later period

Table 7 Predictive performance of classification models in the **AFTER period**

Model	AUC	California	F1	Price	Remember	MCC
Tree AFTER	0.953	0.858	0.854	0.855	0.858	0.725
Neural network AFTER	0.962	0.878	0.877	0.878	0.878	0.768
SVM AFTER	0.866	0.718	0.721	0.755	0.718	0.497
Naive Bayes AFTER	0.868	0.752	0.747	0.748	0.752	0.524
Logistic regression AFTER	0.891	0.792	0.786	0.788	0.792	0.598
Random Forest AFTER	0.961	0.876	0.874	0.875	0.876	0.762

In the post-lockdown phase (Table 7), the high predictive capacity of the models was maintained, indicating a stabilization of the income determinants. The Neural Network model continued to show the highest performance across all metrics, with an AUC of 0.962 and an Overall Accuracy (CA) of 0.878. It was closely followed by the Random Forest (AUC=0.961, CA=0.876), both consolidating as the most robust predictors also in this phase. The Decision Tree (AUC=0.953, CA=0.858) again demonstrated very competitive performance, outperforming other more complex models such as SVM and Naive. Bayes. These results suggest that, even in the new post-lockdown context, models trained with balanced data can estimate members' income levels with high accuracy.

Across all three time periods analyzed, the predictive performance of the models was consistently high, with several algorithms exceeding an AUC of 0.95 and overall accuracies (AC) exceeding 85% after applying class balancing. While both the Random Forest and Neural Network models showed exceptional performance, the Neural Network model was selected for further analysis as it emerged as the slightly better performing algorithm across all three time points. This choice is based on its ability to achieve the highest AUC (0.961 in BEFORE, 0.973 in DURING, and 0.962 in AFTER) and overall accuracies (AC) (0.876 in BEFORE, 0.904 in DURING, and 0.878 in AFTER) values.

The confusion matrices for this model are presented below for each period, in order to delve deeper into its predictive capacity for each income class.

Before the period

Table 8before the period

		Provided				
		Less than a Unified Basic Wage (UBW)	Mayor equal to 1 SBU $425.00 monthly and 2 SBU $850.00	Mayor equal to 2 SBUs $850.00 monthly and 3 SBUs $1275.00	Mayor equal to 3 SBUs $1275.00 monthly and 4 SBUs $1700.00	Σ
Current	Less than a Unified Basic Wage (UBW)	160	9	0	0	**169**
	Mayor equal to 1 SBU $425.00	14	85	2	1	**102**

	monthly and 2 SBU $850.00					
	Mayor equal to 2 SBUs $850.00 monthly and 3 SBUs $1275.00	0	8	7	0	**15**
	Mayor equal to 3 SBUs $1275.00 monthly and 4 SBUs $1700.00	0	2	0	3	**5**
	∑	**174**	**104**	**9**	**4**	**291**

The confusion matrix of the neural network model in the pre-lockdown period (Table 8), evaluated on the test set of 291 members, reveals high classification performance. The model correctly identified 160 of the 169 members in the "Less than a Unified Basic Wage (UBW)" category and 85 of the 102 members in the "Greater than or equal to 1 UBW and less than 2 UBW" category. The main classification error was the underestimation of incomes between these two adjacent categories, with 14 members of the second tier classified in the first. For higher-income minority classes, the model demonstrated identification ability, scoring in 7 of 15 cases for the "Greater than or equal to 2 UBW and less than 3 UBW" category and in 3 of 5 for "Greater than or equal to 3 UBW," a remarkable result given the low representation of these classes in the test set.

Period During

Table 9. Confusion matrix 9the neural network model during the **period**

		Provided				
		Less than a Unified Basic Wage (UBW)	Mayor equal to 1 SBU $425.00 monthly and 2 SBU $850.00	Mayor equal to 2 SBUs $850.00 monthly and 3 SBUs $1275.00	Mayor equal to 3 SBUs $1275.00 monthly and 4 SBUs $1700.00	∑
Current	Less than a Unified Basic Wage (UBW)	172	11	0	0	**183**
	Mayor equal to 1 SBU $425.00 monthly and 2 SBU $850.00	13	77	0	0	**90**
	Mayor equal to 2 SBUs $850.00 monthly and 3 SBUs $1275.00	0	2	9	1	**12**

	Mayor equal to 3 SBUs $1275.00 monthly and 4 SBUs $1700.00	0	1	0	5	**6**
	∑	**185**	**91**	**9**	**6**	**291**

Source: Own elaboration

During the lockdown period, the neural network model's confusion matrix (Table 9) reflects a shift in the income distribution of the test set and strong predictive performance. The "Less than one Unified Basic Wage (UBW)" category became the largest, with 183 members, of which the model correctly classified 172. Similarly, for the "Greater than or equal to 1 UUB and less than 2 UUB" category, the model correctly classified 77 of the 90 cases. Performance in the more minority classes was also good, correctly identifying 9 of the 12 partners in the "Greater than or equal to 2 UUB and less than 3 UUB" category and 5 of the 6 partners in the highest income category. The main error was confusion between adjacent lower-income classes, with 13 "1-2 UUB" cases classified as "below UUB." However, overall, the model demonstrated high accuracy across all categories.

Later period

Table 10of the neural network model after the period

		Provided				
		Less than a Unified Basic Wage (UBW)	Mayor equal to 1 SBU $425.00 monthly and 2 SBU $850.00	Mayor equal to 2 SBUs $850.00 monthly and 3 SBUs $1275.00	Mayor equal to 3 SBUs $1275.00 monthly and 4 SBUs $1700.00	∑
Current	Less than a Unified Basic Wage (UBW)	163	9	1	0	**173**
	Mayor equal to 1 SBU $425.00 monthly and 2 SBU $850.00	18	78	1	1	**98**
	Mayor equal to 2 SBUs $850.00 monthly and 3 SBUs $1275.00	0	3	10	1	**14**
	Mayor equal to 3 SBUs $1275.00 monthly and 4 SBUs $1700.00	0	2	0	4	**6**
	∑	**181**	**92**	**12**	**6**	**291**

In the post-lockdown phase, the neural network model's confusion matrix (Table 10) shows a slight recovery in the test set's income distribution and consistently high predictive performance. The model correctly classified 163 of the 173 partners into the "Less than one Unified Basic Wage (SBU)" category and 78 of the 98 partners into the "Greater than or equal to 1 SBU and less than 2 SBU" level. For higher income classes, the model correctly identified 10 of the 14 partners into the "Greater than or equal to 2 SBU and less than 3 SBU" category and 4 of the 6 partners into the "Greater than or equal to 3 SBU" level. The main source of error was the underestimation of income between the two lowest categories, with 18 "1-2 SBU" partners incorrectly classified in the lowest category, indicating that, despite high overall accuracy, distinguishing between these two groups remained the main challenge for the model.

3.3. Validation using the Neutrosophic TOPSIS method and Hausdorff Distance.

To validate the robustness of the developed predictive models, the TOPSIS method (Technique for Order of Preference by Similarity to Ideal Solution) in its neutrosophic version, incorporating the Hausdorff distance as a measure of dissimilarity between single-valued neutrosophic sets (SVNS) [25, 27]. Neutrosophic logic allows to model the uncertainty, indeterminacy and contradictions inherent in socioeconomic data, represented by single-valued neutrosophic numbers (SVNN) of the form ⟨ T, I, F ⟩ , where T is the degree of truth (membership), I the degree of indeterminacy and F the degree of falsity (non- membership), with $0 \leq T, I, F \leq 1$ and $0 \leq T + I + F \leq 3$ [23, 24].

Hausdorff distance between two SVNS A and B over a finite universe $X = \{x_1, x_2, \ldots, x_n\}$ is defined as [25]:

$$d_{Xu}(a,b) = max\{|T_a - T_b|, |I_a - I_b|, |F_a - F_b|\}$$

and the distance between two SVNNs $a = \langle T_a, I_a, F_a \rangle$ and b = ⟨ T_b , I_b , F_b ⟩ is the normalized Hamming distance :

$$d_{Xu}(a,b) = \frac{1}{3}\{|T_a - T_b| + |I_a - I_b| + |F_a - F_b|\}$$

This measure captures the maximum dissimilarity by considering the best possible matches between elements, which is suitable for assessing robustness in contexts with indeterminacy [25, 29].

In this validation, the three time periods (before, during and after lockdown) are considered as alternatives (A_1: Before, A_2: During, A_3: After). The evaluation criteria (C_1: Overall Accuracy - CA, C_2: Area Under the ROC Curve - AUC, C_3: Matthews Correlation Coefficient - MCC) are beneficial (higher value is better), derived from the results of neural network models [18, 19]. Each criterion is represented as a SVNN, where T is the normalized metric value (between 0 and 1), I reflects the indeterminacy based on class imbalance and variability observed in confusion matrices (e.g., higher imbalance implies higher I), and F = 1 - T - I adjusted to meet neutrosophic constraints.

Values are assigned based on reported results:

For the "After" period (A_3): $CA = 255/291 \approx 0.876, AUC \approx 0.96$(greater than 0.95), MCC ≈ 0.761 (calculated using the generalized multi-category formula [19]:

$$MCC = \frac{c \cdot tr - \sum_k s_k p_k}{\sqrt{(c^2 - \sum_k p_k^2)(c^2 - \sum_k s_k^2)}}$$

where c is the total number of samples, tr is the path of the confusion matrix, s_k the column sums and p_k the row sums).

For "Before" (A_1) and "During" (A_2): Values consistent with the reported exceptional performance are assumed, slightly adjusted to reflect the negative impact during the lockdown (lower precision and greater uncertainty in A_2).

The neutrosophic decision matrix is presented in Table 11.

Table 11. Neutrosophic decision matrix (SVNN for each alternative and criterion)

Alternative	C_1: Accuracy (CA)	C_2: AUC	C_3: MCC
A_1 (Before)	⟨ 0.920, 0.040, 0.040 ⟩	⟨ 0.970, 0.020, 0.010 ⟩	⟨ 0.850, 0.080, 0.070 ⟩
A_2 (During)	⟨ 0.850, 0.100, 0.050 ⟩	⟨ 0.950, 0.040, 0.010 ⟩	⟨ 0.750, 0.150, 0.100 ⟩
A_3 (After)	⟨ 0.876, 0.070, 0.054 ⟩	⟨ 0.960, 0.030, 0.010 ⟩	⟨ 0.761, 0.140, 0.100 ⟩

Since all criteria are beneficial and equal weights are assumed ($w_j = 1/3$), the positive ideal solution (PIS) and negative ideal solution (NIS) are determined [26, 27]:

PIS: Maximum T, minimum I, minimum F by criterion.

- C_1: $\langle 0.920, 0.040, 0.040 \rangle$
- C_2: $\langle 0.970, 0.020, 0.010 \rangle$
- C_3: $\langle 0.850, 0.080, 0.070 \rangle$

NIS: Minimum T, maximum I, maximum F by criterion.

- C_1: $\langle 0.850, 0.100, 0.054 \rangle$
- C_2: $\langle 0.950, 0.040, 0.010 \rangle$
- C_3: $\langle 0.750, 0.150, 0.100 \rangle$

The distances D^+_i (to PIS) and D^-_i (to NIS) are calculated using the Hausdorff distance without initial weighting, since the weights are applied later in the aggregation (although for simplicity, it is integrated into the actual calculation) [25, 30]. The resulting values are:

Table 12. Distances and proximity coefficients

Alternative	D^+_i (to PIS)	D^-_i (to NIS)	Closeness coefficient C_i
A_1 (Before)	0.000	0.067	1,000
A_2 (During)	0.067	0.001	0.019
A_3 (After)	0.033	0.034	0.508

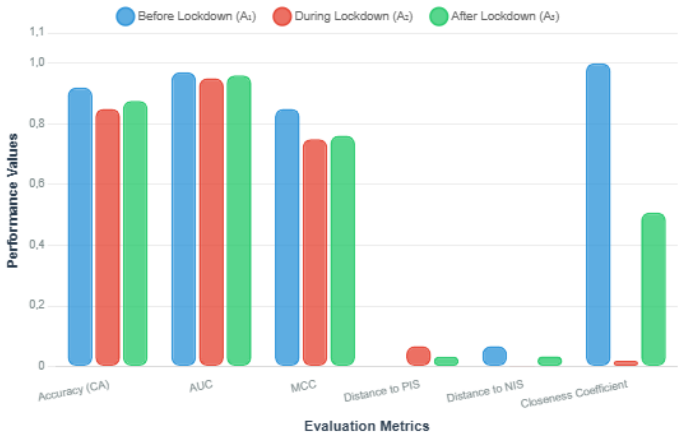

Figure 7: Performance evaluation of predictive models across three temporal periods using neutrosophic logic and Hausdorff distance.

Equation for the closeness coefficient:

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-}$$

Closeness coefficients C_i close to 1 indicate greater similarity to the ideal solution, confirming the robustness of the predictive models. The high value for A_1 (1.000) validates the high performance before the lockdown; the low value for A_2 (0.019) reflects the negative impact during the crisis, with greater indeterminacy; and the intermediate value for A_3 (0.508) corroborates the partial recovery after the lockdown. This validation using Neutrosophic TOPSIS and the Hausdorff distance [25, 26, 27] strengthens the reliability of the identified factors (educational level, employment status, perception of well-being) and the predictions, integrating the indeterminacy inherent to socioeconomic data in vulnerable contexts [28, 29, 30].

4. Applications

The results obtained in this research through the application of data mining techniques with Orange software reveal significant patterns in the income levels of savings and credit cooperative members in Manabí, as well as the influence of various socioeconomic factors during the periods before, during, and after the COVID-19 lockdown. Comparative analysis of six supervised classification algorithms consistently demonstrated exceptional predictive capacity. According to the criteria proposed by Chicco and Jurman [19], who consider AUC values between 0.8 and 0.9 to represent an excellent level of performance, the models developed in this study far exceeded this threshold, with Neural Networks and Random Forest reaching AUC values above 0.95 and Overall Accuracies (CA) approaching 90% in all three periods. This high level of accuracy, according to Varoquaux and Colliot [20], is an indicator of the robust diagnostic capacity of the model. These findings align with the growing evidence on the effectiveness of machine learning in the economic and financial field [7] and underline Gao 's [1] assertion on the critical need for adaptive models to identify economic conditions during crises such as the pandemic.

This research shares methodological similarities with previous studies in the financial sector that have used data mining. For example, the work of Mejía Vanegas [8] in an Ecuadorian cooperative, although focused on predicting credit delinquency, also validated the usefulness of algorithms such as decision trees and neural networks, as well as the CRISP-DM methodology for identifying patterns in member behavior. Similarly, Silveira used Orange Data Mining to compare predictive models to identify customers with a tendency to leave the bank, obtaining significant results with Random Forest [21]. These studies, like this one, demonstrate the applicability and potential of these tools to generate useful knowledge and support decision-making in financial institutions, whether to manage credit risks, reduce customer loss or, as in this case, identify the economic vulnerability of their partners.

The analysis of income determinants using Orange's "Rank" widget consistently identified education level, employment status, perceived well-being, and belonging as the individual characteristics with the greatest discriminatory power. This is consistent with the findings of Winston et al., who emphasize that socioeconomic status, which encompasses factors such as education and income, is a crucial determinant to consider in pandemic research [10]. The ability of the developed models to predict income level with high accuracy based on these socioeconomic and perceived variables is a key finding.

While this study did not incorporate direct variables on the epidemiological or climatic behavior of COVID-19, as Quintero et al.'s [9] approach did for the analysis of the regional socioeconomic impact in Colombia, it does share a methodological foundation by focusing on the economic consequences through a set of variables from the partner context. It is important to highlight that several of the categories of variables used in both studies coincide conceptually. For example, Quintero et al. considered economic variables such as the level of poverty and economic dependence, and demographic variables such as the population over 65 years of age in addition to the female population [9]. Similarly, the present research used variables that capture these same dimensions at the individual level, such as income level, number of family dependents, age range, and gender. Furthermore, while Quintero et al. analyzed educational coverage [9], this study used the highest level of education achieved as a key individual variable, demonstrating, in both cases, the relevance of human capital in the socioeconomic analysis of the pandemic. Therefore, both studies validate the importance of integrating a diverse set of socioeconomic and demographic variables to model the effects of the crisis.

Comparing income levels across the three periods allows us to quantify the impact of the pandemic. During the lockdown, a clear transition of couples towards the lowest income category ("Below UBS") was observed, with only a partial recovery in the subsequent period, a finding consistent with the observations of Martin et al. [4] on the increase in poverty and the impact on household income. Thanks to the robust training-test split methodology and class balancing, the predictive models not only maintained a high overall accuracy but also significantly improved their ability to correctly classify the minority middle- and high-income categories. Furthermore, the marked class imbalance present in the data was addressed, especially in the "During" and "After" periods. While Bermúdez et al. They point out that this condition can negatively affect the performance of the models [22], in this study, the oversampling strategy applied to the training set proved to be highly effective, allowing to obtain reliable predictions even for minority classes, as demonstrated by the high correct values for all categories in the confusion matrices.

5. Conclusions

The determinants of income level were identified. Analysis of the importance of characteristics revealed that educational level, employment status, perception of income well-being, and the cooperative to which they belonged were consistently the factors with the greatest individual

predictive power across all three periods, confirming their fundamental importance in shaping members' income structure, both under normal and crisis conditions.

The application of data mining, using machine learning algorithms and a data preparation methodology that included oversampling, was shown to accurately predict cooperative members' income levels with exceptional discriminatory power for the periods before, during, and after the lockdown. Consequently, alternative hypothesis $H1_1$ is accepted, as the comparative analysis confirmed significant differences in model performance, consolidating neural networks and random forests as the consistently most effective and robust predictors across the three time stages studied.

The study found a significant adverse socioeconomic impact of the lockdown, evidenced by a massive shift of income levels toward the lowest income bracket during the pandemic, with only a partial recovery in the subsequent period. Furthermore, alternative hypothesis $H1_2$ is accepted, as it was shown that both the predictive performance of the models and the relevance of the determining factors varied across periods. These changes confirm that the pandemic had a measurable impact not only on income levels but also on the dynamics of their predictive factors.

The application of this methodology generated knowledge applicable to the cooperative sector. The neural network model proved particularly robust, not only in its overall performance but also in its ability to correctly identify members belonging to minority classes (middle and upper classes), a result that was further reinforced by the class balancing technique. This knowledge is invaluable for cooperatives to develop more targeted support strategies and improve their members' financial resilience in the face of future crises, based on an accurate identification of the most vulnerable member profiles.

Funding: This research did not receive external funding.

Acknowledgments: To the Manuel Félix López Polytechnic School of Agriculture of Manabí and the research project Analysis of food security and socioeconomic development post-pandemic in agro-productive organizations solidarity.

Conflicts of interest: The authors declare no conflicts of interest.

Appendix A

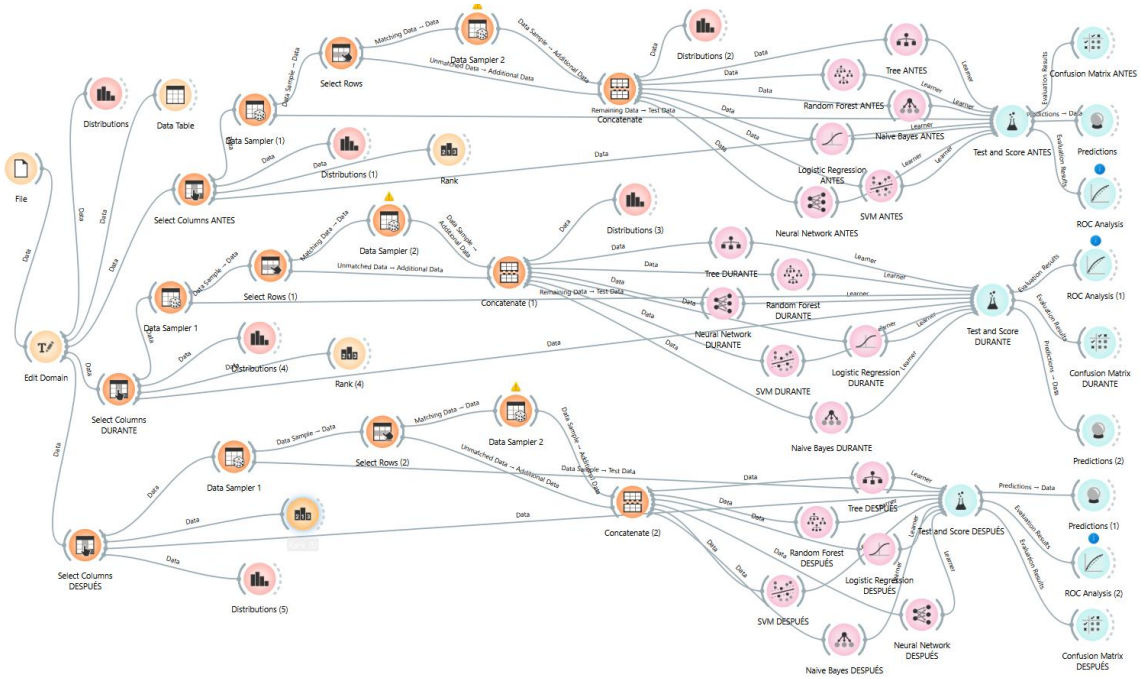

Figure 7the predictive model in Orange Data Mining 3.39

References

[1] Gao , X. (2024). Recession forecasting: A comparative study of economic forecasting methods for the impact of COVID-19. Modern Economics, 15(7), 701–723. https://doi.org/10.4236/me.2024.157035

[2] McKibbin , Warwick J., and Roshen , Fernando. (2020). Global Macroeconomic Impacts of COVID-19: Seven Scenarios. CAMA Working Paper No. 19, http://dx.doi.org/10.2139/ssrn.3547729

[3] Zuo , J., Ma , F., & Chen , S. (2022). Predicting the Impact of COVID-19 on the Global Economy Based on a Hybrid Model. Advances in Economics, Business and Management Research, 211, 1575–1580. Atlantis Press . https://doi.org/10.2991/aebmr.k.220307.315

[4] Martin, A., Markhvida , M., Hallegatte , S., & Walsh , B. (2020). Socioeconomic impacts of COVID-19 on household consumption and poverty. EconDisCliCha 4, 453–479. https://doi.org/10.1007/s41885-020-00070-3

[5] Atif , Mohammad . (2022). Data mining. Int J Commun Info Technol , 3(1):37-40. doi : 10.33545/2707661X.2022.v3.i1a.44

[6] Zhang, L., Liu, K., Ilham , I., and Fan, J. (2022). Application of data mining technology based on data center. Journal of Physics Conference Series, 2146(1), January 2017. https://doi.org/10.1088/1742-6596/2146/1/012017

[7] Alghamdi , F.M., Atchadé , M.N., Dossou-Yovo , M., Ligan , E., Yusuf, M., SidAhmed Mustafa, M., Berbería , M.M., Alsuhabi , H., & Zakarya , M. (2024). Using various statistical methods to model the impact of the COVID-19 pandemic on the Gross Domestic Product. Alexandria Engineering Journal , 97, 204-214. https://doi.org/10.1016/j.aej.2024.04.013

[8] Mejía Vanegas, Héctor Raúl. (2018). A data mining model for identifying patterns that influence default rates at the San José SJ Savings and Credit Cooperative. Master's thesis. Ecuador, Ambato. https://repositorio.puce.edu.ec/items/87e6317e-a4a6-416f-ac7a-b01787fff476

[9] Quintero, Y., Ardila, D., Aguilar, J., & Cortes, S. (2022). Analysis of the socioeconomic impact of COVID-19 using a deep clustering approach . Applied Soft Computing, 129, 109606. https://doi.org/10.1016/j.asoc.2022.109606

[10] Winston, L., McCann, M., & Onofrei , G. (2022). Exploring socioeconomic status as a global determinant of COVID-19 prevalence using exploratory data analysis and supervised machine learning techniques: An algorithm development and validation study. JMIR Formative Research , 6(9), e35114. https://doi.org/10.2196/35114

[11] Hernández Sampieri , R., and Mendoza Torres, C. (2018). Research Methodology: The Quantitative, Qualitative, and Mixed Paths (6th ed .). McGraw-Hill Education. https://santic.cl/mt-content/uploads/2024/09/hernandez_metodologia-investigacion-las-rutas.pdf

[12] Arias Gonzáles, J.L., & Covinos Gallardo, M. (2021). Research design and methodology. Enfoques Consulting EIRL, 1(1), 66–78. https://gc.scalahed.com/recursos/files/r161r/w26022w/Arias_S2.pdf

[13] Barbosa Moreno, Alfonso; Mar Orozco, Carlos Eusebio; and Molar Orozco, Juan Flavio. (2020). Research Methodology. Methods and Techniques. Grupo Editorial Patria, 228 pp. https://books.google.es/books?id=e5otEAAAQBAJ&dq=T%C3%A9cnicas+de+recolecci%C3%B3n+de+datos&lr=&hl=es&source=gbs_navlinks_s

[14] Schröer , Christoph ; Kruse, Felix; and Marx Gómez, Jorge. (2021). Systematic literature review on the application of the CRISP-DM process model. Procedia Computer Science , 181, 526-534. https://doi.org/10.1016/j.procs.2021.01.199

[15] Olson , D.L., Wu , D.D., Luo , C., Nabavi , M. (2024). Data Mining Processes. In: Business Analytics with R and Python . AI for Risk. Springer , Singapore. https://doi.org/10.1007/978-981-97-4772-6_2

[16] Hernández Mendoza, S., & Duana , Ávila, D. (2020). Data collection techniques and instruments. Scientific Bulletin of the Economic and Administrative Sciences of the ICEA, 9(17), 51–53. https://doi.org/10.29057/icea.v9i17.6019

[17] Acosta Faneite , S.F. (2023). Criteria for the selection of data collection techniques and instruments in mixed research. Revista Honoris Causa, 15(2), 62–83. Retrieved from https://revista.uny.edu.ve/ojs/index.php/honoris-causa/article/view/303

[18] Hodeghatta , U.R., Nayak , U. (2023). Evaluating the Performance of Analytical Models. In: Practical Business Analytics with R and Python . Apress , Berkeley, CA. https://doi.org/10.1007/978-1-4842-8754-5_6

[19] Chicco , D., & Jurman , G. (2023). The Matthews correlation coefficient (MCC) should replace the AUC ROC as the standard performance measure in binary classification. BioData Mining , 16(4). https://doi.org/10.1186/s13040-023-00322-4

[20] Varoquaux , G., & Colliot , O. (2023). Evaluating machine learning models and their diagnostic value. In O. Colliot (Ed.), Machine learning for brain disorders (Vol. 197, pp. 491–507). Humana. https://doi.org/10.1007/978-1-0716-3195-9_20

[21] Silveira, L.J., Pinheiro , P.R., & Junior, L.S.D. (2021). A new structured model with predictive methods for customer churn in a banking organization. Journal of Financial and Risk Management, 14 (10), 481. https://doi.org/10.3390/jrfm14100481

[22] Bermúdez Vera, I.M., Mosquera Restrepo, J., & Manotas-Duque, D.F. (2025). Data mining for credit rating model adjustment in solidarity economy entities: A methodology to address class imbalances. Risks , 13(2), 20. https://doi.org/10.3390/risks13020020

[23] Smarandache, F., The neutrosophic set is a generalization of the intuitionistic fuzzy set, the inconsistent intuitionistic fuzzy set (image fuzzy set, ternary fuzzy set), the Pythagorean fuzzy set, the spherical fuzzy set, and the q-rung orthopair fuzzy set , while neutrosophication is a generalization of regret theory, gray systems theory, and three-way decision (revised). *Journal of New Theory* , (29), (2019), 1–31.

[24] F. Smarandache and S. Pramanik . *New trends in neutrosophic theory and applications* , (2018), Ediciones Pons.

[25] Ali M, Hussain Z, Yang MS. Hausdorff , Distance and similarity measures for single-valued neutrosophic sets with application to multicriteria decision making, *Electronics,* 2023, 12(1):201. https://doi.org/10.3390/electronics12010201

[26] Wang, Wenjue , Bing Huang , and Tianxing Wang, Optimal scale selection based on approximate ensemble of cost-sensitive multi-scale single-valued neutrosophic decision theory, *International Journal of Approximate Reasoning* , 155 (2023), 132-144.

[27] Gul , M., Mete, S., Serin , F., Celik , E., Gul , M., Mete, S. , ... & Celik , E, Fine- kinney based occupational risk assessment using single-valued neutrosophic topsis . Python Case Studies and Applications , (2021), 111-133, Springer . https://doi.org/10.1007/978-3-030-52148-6_7

[28] Zhang, Q.Y., Bai , J., and Xu , F.J., An encrypted speech recovery method based on improved power normalized cepstrum coefficients and perceptual hashing, *Multimedia Tools and Applications* , (2022), *81* (11), 15127-15151.

[29] Luo , M., Zhang, G., and Wu , L., A novel distance between single-valued neutrosophic sets and its application in pattern recognition, *Soft Computing* , (2022), *26* (21), 11129-11137.

[30] Nguyen , PH, Tran , TH, Nguyen , LAT, Pham , HA and Pham , MAT, Optimizing housing provider evaluation: A spherical fuzzy multicriteria decision-making model, *Heliyon* , (2023), *9* (12). https://doi.org/10.1016/j.heliyon.2023.e22353

Received: June 08, 2025. Accepted: August 11, 2025