Use of a Neural Network to Measure the Impact of Social Distribution and Access to Infrastructure on the HDI of the Municipalities of Mexico

Fernando I. Becerra López, Ricardo Pérez Ramírez

Department of Mathematics, Universidad de Guadalajara, Guadalajara, Mexico
Email: ignacio.becerra@academicos.udg.mx, ricardo.perez8864@alumnos.udg.mx

Abstract

The Human Development Index (HDI) was created by the United Nations (UN) and is the basis for many other indicators, as well as being the origin of many public policies worldwide. It is a summary measure of life expectancy, education, and per capita income. These components, in addition to being global measures, show difficulty in being impacted and, with this, advancing in the level of human development. This work shows a model that relates variables of social distribution and access to infrastructure in Mexico, with the HDI. These variables were chosen through a statistical analysis based on a set of indicators measured by the National Institute of Statistics and Geography (INEGI) periodically at the municipal level. The statistical analysis shows that there is no simple correlation between these variables and the HDI, so that a supervised learning model based on a neural network was used, therefore proposing a classification technique based on the distribution of data in the underlying metric space. In addition, an attempt was made to find the simplest possible model to reduce the computational cost and in turn obtain information on the variables with the greatest impact on the HDI, with the aim of facilitating the creation of public policies that impact it.

Keywords

Multilayer Perceptron, Human Development Index, K-Means, Non-Linear Correlation

1. Introduction

In all countries there are indices and indicators that help governments monitor

the performance of their policies, these can refer to education, health, infrastructure, and social distribution, among others; although these are only methodological proposals and are likely to receive comments to improve their usefulness, for example, as in [1] where a different way of evaluating marginalization in Mexico is proposed. An advantage of having diverse types of indicators is that with these it is possible to make analyses between different indicators on an objective indicator, as in [2] where development is taken as a variable that is influenced by distinct factors, features such as social, and economic, among others.

With the above in mind, the Human Development Index, HDI, is selected for this work as an index that reflects the quality of life of a population and taking the view that the indices are impacted not only by the methodology with which they are created but also by other features, a selection of other indices are proposed, which have no appreciable direct relationship but at the same time it is inferred that modifications to these have an impact on the quality of life of a population.

HDI was published for the first time in 1990 by the United Nations Development Program, UNDP (1990). This index was introduced due to the need to have a measure of development in the countries and its fundamental objective is to measure the development of the human being, unlike, for example, the Gross Domestic Product, GDP, of a country, which reflects development, but based only on its economic activity. Therefore, three components were chosen to calculate the HDI, which focuses on health, education, and wealth, which represent the fundamental axes in the development of a person. HDI has become a very important tool for governments, including organizations such as the Economic Committee for Latin America and the Caribbean (CEPAL) [3].

The health indicator, calculated by the longevity of a population, is determined by the life expectancy at birth of a person. It is of special relevance since it indirectly reflects a population's access to health services, as well as adequate nutrition, since, without these two indirect characteristics, it would be difficult to increase life expectancy.

Regarding the education component, this is calculated with the literacy rate of a population, something of significant importance since it provides the opportunity to access knowledge. In fact, it is currently desired that the population have access to higher levels of knowledge for a better performance in their productive lives.

Lastly, the index of the wealth of a population tries to reflect the capacity that this must face the basic needs that an individual may have for its development. This index is calculated with the GDP per capita of each country or region together with a correction, purchasing power parity, to homogenize the level of said purchasing power between different regions.

As can be seen, the three components of the HDI really aim to reflect what an individual's life of well-being can be like, a long life with access to education to develop the desired and well-paid economic activity. Unfortunately, these three components are averages in the population and, therefore, a global measure,

which can hide the reality and the dispersion that the population experiences in each of these factors.

In addition to the, designing public policies that help increase these three components and, consequently, the HDI does not turn out to be intuitive. Therefore, the search for other indices with a more sensitive impact on the decisions made by the government can be helpful in the design and implementation of public policies that help improve the HDI of the regions.

2. Variables Proposed to Influence the HDI Level

Considering what was expressed in the introduction, indices were selected that are believed to be more local and easily obtained (all are provided by the National Institute of Statistics and Geography, INEGI, from the year 2010, and at the municipal level, where the methodological manuals are in [4] and [5]). These indices or variables are the following.

The percentage of the population that lives in communities of less than 5,000 inhabitants in a municipality (PL < 5000), is a variable that is used in the reports of marginalization prepared by the Government of Mexico, which is important, since in more than 50% of the municipalities in the country have 100% in this index, in addition populations of this type tend to have less access to services and less economic development.

The Labor Force Participation Rate (LFPR) of a municipality, which refers to the quotient of economically active people who are working or looking for a job (a person can conduct an economic activity from the age of fifteen) between the entire population.

$$LFPR = \frac{LF_{(15 \text{ or more})}}{P_{(15 \text{ or more})}} \times 100 \tag{1}$$

where: $LF_{(15 y \text{ mas})}$ is the labor force aged fifteen or over and $P_{(15 y \text{ mas})}$ is the total population greater than or equal to fifteen years.

The degree of accessibility to paved roads (AccesInfra), which is obtained thanks to the work of the National Council for the Evaluation of Social Development Policy, CONEVAL, and which reflects the ease that different communities have in using paved roads. The AccesInfra grade per municipality is taken as the weighted sum of the AccesInfra's grades per community, which make up the municipality.

$$\frac{\sum_{i=1}^{n} (p_i \times g_i)}{\sum_{i=1}^{n} p_i} \tag{2}$$

where p_i is the population of the i-th community that makes up the municipality, g_i is the degree of accessibility of the i-th community and n is the number of communities that make up the municipality.

In addition, the percentage of the population of a municipality which does not native to the same federal entity (%PobMig) was chosen, considering that the phenomenon of migration is intricately linked to the search for better job op-

portunities and life prospects. Due to this, municipalities that have high scores in this index could be interpreted as municipalities that offer high standards of living and that is why they attract populations from other states.

Lastly, the population density of the municipality (DENS10), which is a variable that might not seem to have as much relevance, but since many of the infrastructure construction decisions are public and private, such as universities, hospitals, etc. is linked to covering the largest possible population, these constructions are aimed at municipalities with densely populated areas.

3. Relationship between the Variables and the HDI

As can be seen, several of the indices presented not only have a municipal focus, but are even obtained from a community level, so that the reality that the inhabitants may be experiencing can be better reflected. Likewise, indices such as the LFPR and the AccesInfra are sensitive to public policies, in addition to the fact that they all function as control variables for human development (although this does not mean that it is impossible to create public policies that help improve them in the short term).

A main objective of this work is to find a relationship between these five variables with the HDI of each municipality, to then indirectly relate to the three axes that support the HDI and facilitate the structuring of public policies based on these five indices and thereby improve the HDI of the municipalities. For example, with the urban planning of services and infrastructure, which better benefits its surrounding communities with a high index of PL < 5000, so that they have greater access than densely populated communities. Also, the search for private investment, national or foreign, for the creation of new jobs, to increase the attraction of population from other states or retention of its own population and that, in turn, would be reflected in the LFPR.

4. The Proposed Model Using a Neural Network

Doing a linear correlation analysis between the selected variables and the HDI, it can be seen that it does not exist for any of them, as can be seen in Table 1, so it will be necessary to use a model that can find non-linear and multivariate correlations, being Neural networks are a good tool for this type of problem, as demonstrated in [6]. For this reason, a Multilayer Perceptron (MLP) was selected as a model [7], which is a generalization of the simple Perceptron proposed by [8], which through its processing units (neurons), and their dynamic states of activation [9], processes the input data in order to find patterns in the data and thereby offer a model capable of generalizing [10].

Before explaining the architecture and the results obtained, it is important to mention that the outputs of an MLP express the probability of belonging to a certain set, therefore, it was decided to classify (cluster) the municipal HDI values into three groups with the method of K-means as in [11]. The decision to classify into three groups was based on a statistical analysis, in which three

Table 1. Correlation matrix.

	AcessInfra	DENS10	PL < 500	LFPR	%PobMig	HDI
AcessInfra	1.0000	0.1778	−0.3870	0.1019	0.2767	0.5617
DENS10	0.1778	1.0000	−0.3398	0.1901	0.3203	0.3299
PL < 5000	−0.3870	−0.3398	1.0000	−0.2044	−0.3990	−0.6309
LFPR	0.1019	0.1901	−0.2044	1.0000	0.1881	0.1942
%PobMig	0.2767	0.3203	−0.3990	0.1881	1.0000	0.4964
HDI	0.5617	0.3299	−0.6309	0.1942	0.4964	1.0000

classes offered greater separability between groups, this against a greater or lesser number of clusters. Once the limits of the classes were obtained, they were labeled, so that the class labeled with [1,0,0] represents the group of municipalities with a "high" HDI, the label [0,1,0] represents the group of municipalities with a "medium" HDI and, finally, the label [0,0,1] is the representative of the municipalities with a "low" HDI.

For the training of the MLP, it was decided to separate the data by municipalities and at the same time validated for robustness by the Student's T test, such that 70% of these were used to train the model and the remaining 30% to validate the model and how capable it is to generalize or, in other words, assess the perdition of membership of an input dataset, on which it was not trained.

The architecture selected for the problem was an input layer with 5 neurons, a hidden layer with 25 neurons, and an output layer with 3 neurons, all with logistic activation function (Equation (3)), trained with the backpropagation algorithm and in an "off-line" mode because the data presented concurrency [12] [13]. This architecture is selected, since a better performance (stability) was observed in the generalization of the data and the norm of its derivative reached almost zero, Figure 1. Also, it was made an analysis of convergence for several amounts of neurons at the hidden layer to guarantee avoid overfitting and keep the model simple.

$$\sigma = \frac{1}{1 + e^{-x}} \tag{3}$$

For the validation of the model, precision was taken as a metric, which is defined by:

$$\text{Acc} = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \times 100 \tag{4}$$

An 81% accuracy was obtained for the training data and 74% for the validation data, giving 79% in the evaluation with all the data, which confirms that there is a correlation of the variables with the HDI of the municipalities.

Already having the model, tests were conducted both to validate it and to observe the behavior of the selected indices. For example, the median of each of the indices was selected, since due to the bias of the data it provides us with a better measure of central tendency, obtaining a prediction of the average HDI, which was expected.
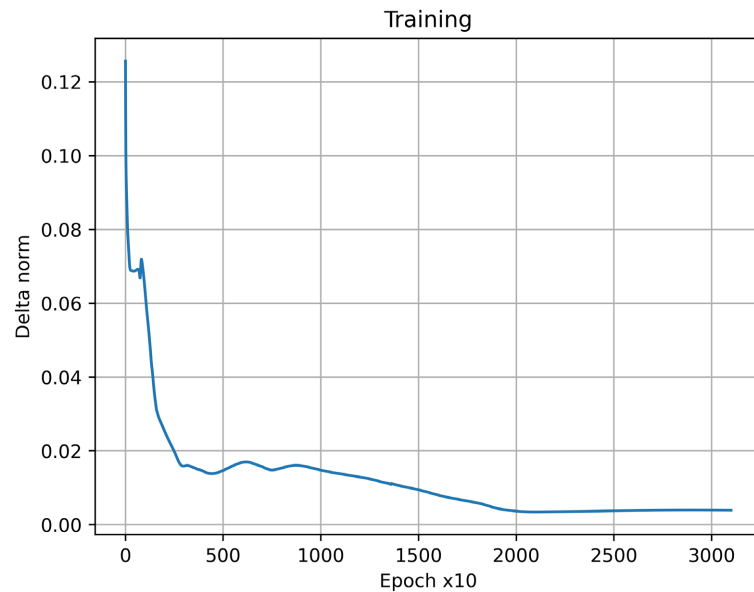
Figure 1. MLP training.

In addition, an arrangement of hypothetical municipalities was created, where each of them had the median in all the indices except one, which is evaluated with its minimum value and then with its maximum value, to observe the relationship between the indices regarding these municipalities. It was observed that the indices with the greatest impact are AccesInfra, %PobMig and PL < 5000, the latter with an inverse relationship to the HDI (that is, a higher value is reflected in a decrease in the HDI). For example, a municipality with all their indices equal to the median, but %PopMig at the minimum obtained a low HDI prediction of 76.75%, while with %PopMig at the maximum a medium HDI prediction of 97.42% was obtained.

After this, the exercise of leaving all the indices at their lowest or highest value was conducted, with only one of them varying between the range of minimum and maximum values. With this, it was identified that only one of the indices, %PobMig, has an impact on the classification made by the model, while the others do not. This tells us, omitting the case, that the movement of a single variable has no relevance and, therefore, there is no one-to-one relationship with the HDI.

Something to confirm the above in a quantitative way and to be able to observe that all the indices are important for the model, an analysis of characteristics was carried out taking the proposal of [14], in which the importance S_i of an input to the model as seen in Equation (5) and being the criterion in [15] the one used. Equation (6), where w_{ij} is the weight of the i-th data to the j-th neuron, thus obtaining **Table 2** of results and where it is observed that the value S of the %PobMig index is the highest and the AccesInfra the lowest, although the latter is still important in scale with the others.

$$S_i = \sum_{j=1}^{n} s_{ij} \tag{5}$$

Table 2. Individual importance of features for the model.

	AcessInfra	DENS10	PL < 500	TPE	%PobMig
S	7157.90	13650.49	15769.78	13267.01	24340.26

$$s_{ij} = \left(w_{ij} \right)^2 \tag{6}$$

This leads us to conclude that the proposed model (MLP) establishes a better relationship between the proposed variables and the HDI over other models such as, for example, multiple linear regression.

5. Conclusions

After observing that the model managed to relate the selected indices and the HDI of the municipalities, a statistical analysis of the municipalities with high HDI is conducted, calculating the averages of the indices, and comparing them with those obtained globally, as well as from the municipalities that are not included in the high HDI cluster.

When conducting the, it is observed that the municipalities with a high HDI have higher average values in access to paved roads, a higher population density, a slightly higher economic participation by their population, higher percentages of migrant population from other states, as well as a smaller percentage of the population living in small communities. Something that is also observed in the relationships found with this model was that AccesInfra, %PobMig and PL < 5000 have greater weight compared to DENS10 and TPE. This suggests that policies focused on improving AccessInfra, %PobMig and PL < 5000, would have a greater positive impact on the HDI of these communities. This is a special approach since public policy decisions are rarely made using machine learning models. Furthermore, the idea of investigating indirect variables and their influence on human development is also new.

An example of this type of public policy for AccesInfra is the policy of the current Federal Government of Mexico to pave access to municipal capitals for municipalities that did not have this. For %PobMig, one can take what has been done in China with its special economic zones that have attracted people from the interior of the country, where the HDI is usually lower than in said special zones. Finally, the PL$ < $5000 is an index that cannot be impacted so quickly, since it will depend on the resources and services that these communities receive to help in their development, urbanization and growth (in the worst case that these small communities disappear and are grouped into a main one in the same municipality or in another part of the country).

For all the above, it can be inferred that creating public policies that consider the selected variables of social distribution and access to paved roads, would have a positive impact on the HDI of the municipalities and, therefore, on the original variables that they calculate it: health, schooling, and GDP.

It is important to note that, although the model performs well with the present

data, it is difficult to interpret the real effects due to the public policies implemented. However, changes to the HDI locally may be reviewed in the future.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Gutiérrez-Pulido, H. and Gama-Hernández, V. (2010) Limitantes de los índices de marginación de Conapo y propuesta para evaluar la marginación municipal en México. *Papeles de población*, **16**, 227-257.

[2] Peláez-Herreros, Ó. (2012) Análisis de los indicadores de desarrollo humano, marginación, rezago social y pobreza en los municipios de Chiapas a partir de una perspectiva demográfica. *Economía, sociedad y territorio*, **XII**, 181-213. https://doi.org/10.22136/est00201290

[3] Salas-Bourgoin, M. (2014) A proposal for a modified Human Development Index. *CEPAL Review*, **112**, 29-44.

[4] Coneval, C.N. (2018) Grado de accesibilidad a carretera pavimentada. Coneval, Ciudad de México. https://www.coneval.org.mx/Medicion/Paginas/Grado_accesibilidad_carretera.aspx

[5] Inegi, I.N. (2017) Metodología de Indicadores de la Serie Histórica Censal. Inegi, Ciudad de México. https://www.inegi.org.mx/contenidos/programas/ccpv/cpvsh/doc/serie_historica_censal_met_indicadores.pdf

[6] Abdulsalama, K.A. and Babatunde, O.M. (2019) Electrical Energy Demand Forecasting Model Using Artificial Neural Network: A Case Study of Lagos State Nigeria. *International Journal of Data and Network Science*, **3**, 305-322. https://doi.org/10.5267/j.ijdns.2019.5.002

[7] Hilera González, J.R. and Martínez Hernando, V.J. (1995) Redes Neuronales Artificiales: Fundamentos, Modelos y Aplicaciones. RA-MA, Madrid.

[8] Rosenblatt, F. (1958) The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, **65**, 386-408. https://doi.org/10.1037/h0042519

[9] Cybenko, G. (1989) Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals, and Systems* (*MCSS*), **2**, 303-314. https://doi.org/10.1007/BF02551274

[10] Lecun, Y. (1989) Generalization and Network Design Strategies. In: Pfeifer, R., Schreter, Z., Fogelman, F. and Steels, L., Eds., *Connectionism in Perspective Elsevier*, Elsevier, Toronto.

[11] Fix, E. and Hodges, J.L. (1989) Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review*, **57**, 238-247. https://doi.org/10.2307/1403797

[12] Møller, M.F. (1993) A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. *Neural Networks*, **6**, 525-533 https://doi.org/10.1016/S0893-6080(05)80056-5

[13] Higham, C.F. and Higham, D.J. (2018) Deep Learning: An Introduction for Applied Mathematicians. arxiv:1801.05894. https://arxiv.org/abs/1801.05894

[14] Sohangir, S. and Gupta, B. (2014) Neuro Evolutionary Feature Selection Using NEAT. *Journal of Software Engineering and Aplications*, **7**, 562-570. https://doi.org/10.4236/jsea.2014.77052

[15] Belue, L.M. and Bauer, K.W. (1995) Determining Input Features for Multilayer Perceptrons. *Neurocomputing*, **7**, 111-121. https://doi.org/10.1016/0925-2312(94)E0053-T