ISSN 2007-9737

# Data Mining Approach for the Prediction of Hypertension and its Correlation with Socioeconomic Factors in Mexico: A Case Study

Obdulia Pichardo-Lagunas[1], Bella Martinez-Seis[1,*], Sabino Miranda[2]

[1] Instituto Politécnico Nacional,
Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas, Mexico City,
Mexico

[2] Universidad Autónoma de la Ciudad de México, Mexico City,
Mexico

{opichardola@ipn.mx, bcmartinez@ipn.mx}@mixteco.utm.mx, smiranda@ieee.org

**Abstract.** According to the Secretary of Health, High Blood Pressure (HBP) has remained among the top ten leading causes of death in Mexico. In recent years, data-driven analysis studies have become a common complement to health research. Therefore, reliable records on this subject are necessary. This paper shows the collection, selection, and integration of a unified database created from different public access sources regarding HBP. We propose a methodology for the identification of correlations between non-medical factors and HBP using data mining techniques such as clustering, we validate them with Pearson Correlation Coefficient. We also used statistical and artificial intelligence models to predict the number of cases of HBP, we evaluated them with Root Mean Square Error and Mean Absolute Percentage Error, the best results were with Convolutional Neural Network Quantile Regression. All in order to generate tools that support the prevention of the future development of hypertension.

**Keywords.**Correlation, data mining, data exploration, hypertension, neural networks.

## 1 Introduction

Non-communicable diseases (NCDs) are not spread through infection or other people but are typically caused by unhealthy behaviors. In the last two decades, there has been an increase in the incidence of NCDs in most countries of the world. According to the World Health Organization (WHO), 70% of the 56.4 million deaths in 2016 were caused by NCDs. Furthermore, 75% of NCD-associated deaths occurred in low- and middle-income countries, reflecting the seriousness of the problem in countries such as Mexico [21]. Among NCDs, chronic vascular diseases are the leading cause of death in the world, including Systemic Arterial Hypertension (SAH), Ischemic Heart Disease, Heart Failure, Degenerative Calcific Aortic Valve Stenosis, and Congenital Heart Disease [21].

Among these diseases, SAH has shown an accelerated epidemiological change [22]. In Mexico, approximately 450 thousand new cases of SAH are diagnosed annually.

In the past two decades, it has remained among the first nine causes of death in Mexico, and in the past six years, the mortality rate from this cause has increased by 29.9% [22].

In addition, Mexico has the highest prevalence of arterial hypertension in the world, according to the National Survey of Health and Nutrition 2012.

This paper proposes to collect data from public records provided by official sources to concentrate them in a database. The data obtained allowed the generation of a predictive model to analyze the behavior of the disease (SAH) in a period of one year using artificial intelligence tools.

Through the use of data mining techniques, the system can also establish a relationship between variables such as socioeconomic status (SES) with SAH. The impact of SES on hypertension has been reported in several studies with conflicting results [15, 17].

doi: 10.13053/CyS-28-4-5215

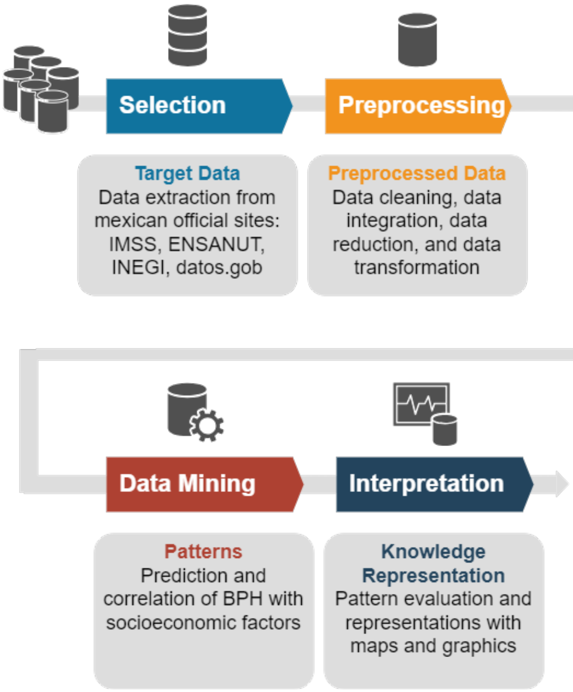
Selection
Preprocessing
Target Data
Data extraction from mexican official sites: IMSS, ENSANUT, INEGI, datos.gob
Preprocessed Data
Data cleaning, data integration, data reduction, and data transformation
Data Mining
Interpretation
Patterns
Prediction and correlation of BPH with socioeconomic factors
Knowledge Representation
Pattern evaluation and representations with maps and graphics

**Fig. 1.** Methodology

**Table 1.** Evaluation of WAPE and RMSE from 1 to 5 years prediction

| Number of Predicted Years | WAPE | RMSE |
|---|---|---|
| 2019 (1) | **0.0825** | **150505.70** |
| 2020 (2) | 0.1029 | 213206.77 |
| 2021 (3) | 0.1517 | 331492.10 |
| 2023 (5) | 0.1736 | 356471.12 |

## 2 Data Analysis and Hypertension

BNP-PL registry is the first multicenter and prospective registry of adult and pediatric patients with PAH and CTEPH created in any of the Central-Eastern European countries [14]. In Greece, a large-scale database of ageing participants was integrated to evaluate the long-term hypertension risk of men and women aged older than 50 years with Machine Learning [8].

In Latin America, Cuba used their data to design a predictive model for the diagnosis of arterial hypertension in adulthood from records taken from adolescence; Perez et al. [18] obtained an effectiveness of 80%, and Gonzalez et al. [2] analyzed the importance of data engineering and obtained a methodology that delivers more effective results 88.41% success in grouping instances. They also conclude that the risk of becoming hypertensive in the trajectory from adolescence to adulthood is mainly due to birth and family factors.

Machine learning classification techniques have been proposed to help healthcare professionals in diagnosing heart disease. Some works focus on the identification of risk patterns associated with hypertension.

In this sense, [19] uses Statistical Tree algorithm, Classification, and Regression, and [20] uses algorithms like K-star, J48, SMO, Naïve Bayes, MLP, Random Forest, Bayes Net, and REPTREE.

It is also important to study the correlation of the disease with other factors. For example with demographic elements [4], Body Mass Index (BMI) [1], race [17], socioeconomic status (SES) [15, 17, 10], or COVID [24]. For this study, we focus on SES.

### 2.1 Analysis of Hypertension in Mexico

The Center for Research in Nutrition and Health and the National Institute of Public Health describes the prevalence, distribution, and trends of arterial hypertension (HTN) in Mexican adults aged 20 years or older, considering 10,898 adults.

The project performed a time trend analysis using statistical analysis based on the data reported in the National Health Survey (2000), the National Health and Nutrition Survey (2006), and the National Health and Nutrition Survey (2012). They found stabilization of the problem between 2000, 2006, and 2012, but better control of the disease is required [5].

A comparative cross-sectional study was carried out in 627 rural communities of Durango, Mexico, and determined blood pressure figures and sociodemographic variables through the IMSS-Solidarity program. For the statistical analysis, the confidence interval was calculated in the binomial distribution for the dichotomous variables (prevalence estimate).

For nominal variables, the $\chi^2$ test was calculated for the significance of the difference between groups, and for continuous variables, the calculation was with the $t$ test. Knowing a possible behavior, progress, or evolution of arterial hypertension has been of interest to students and researchers; they have worked in different ways to detect possible behaviors in the future. These projects have worked with samples that are not greater than 120,028 individuals taken in periods of less than eight years.

**Table 2.** Numer of null and NaN

| DataSet | No. of cells | No. of NaN | No. of NULL |
|---|---|---|---|
| Ensanut (Economy) | 22,487,304 | 224,198 | 3,699,427 |
| Ensanut (Food) | 1,556,334 | 25,476 | 30,915 |
| Ensanut (Activity) | 1,447,800 | 453,171 | 96,923 |
| Ensanut (Teenagers) | 1,261,248 | 0 | 661,073 |
| Ensanut (Adults) | 1,596,425 | 108,310 | 756,868 |
| INEGI (Deaths) | 8,745,373 | 0 | 0 |
| Health Sector (HAS-Obesity) | 37,510,485 | 0 | 14,580,308 |
| IMSSS (HAS) | 756 | 0 | 0 |

**Table 3.** Evaluation of WAPE and RMSE in different models for the prediction of 2019

| Model | WAPE | RMSE |
|---|---|---|
| Mean ARIMA | 0.0825 | 150505.70 |
| CNN-QR | **0.0337** | **40604.83** |

The solution designed in this document works with a larger number of samples collected over a longer time interval and considers using data mining tools.

## 3 Methodology: Data Mining Approach with Hypertension Data

This section describes the different stages of the proposed solution: data collection and analysis, pre-processing, data mining application, and data interpretation. The proposed project includes data on hypertension and socioeconomic indicators. The information on each of these variables was subjected to the following stages: data collection, data selection, data pre-processing, application of data mining, and data interpretation, as we can see in Figure 1.

### 3.1 Data Collection and Selection for HBP

We use data about hypertension and socioeconomic status. For the first one, the Mexican government keeps track of patients with hypertension; and for the second one, data was collected from Instituto Nacional de Estadística y Geografía (INEGI) [11] and Open Data of Mexican government [6]. The original data was presented in multiple formats, with different characteristics, and in some cases present inconsistencies. The data collection period comprises from August 2019 to May 2020. Databases were consulted and collected from the official sites of ENSANUT, IMSS, ISSSTE, INEGI, and the Health Sector.

These data was available to the Mexican population in files with different formats, distributed in portals of each institution. The selection considered mainly the completeness of data. The final data set considers data from the institutions of ENSANUT, IMSS, INEGI, and the Health Sector. A deeper description of the data set in presented in Section 4.

### 3.2 Preprocessing

First, data exploration of the datasets was performed. Then an iterative data cleaning process was executed using data reduction and data transformation in order to generate an integral data set.

Data wrangling was performed, it allowed to transform data into a unified format. This process reduced the dimensionality of each database. Then, we looked for null and empty data that is presented in 4. If a column has more than 30% of null or empty data records, we drop it. Some other inconsistencies as dates and different indicators for the same thing were solved. A unification and transformation of data types was also required.

The preprocessing stage was applied separately to the records corresponding to Hypertension, Nutrition, and Socioeconomic level. Then, integration stage was executed, joining the datasets of the same institution with data transformation in some cases looking for the consistency of the data.

### 3.3 Forecast and Correlation with Socioeconomic Factors

A pattern means that the data are correlated, have a relationship, or are predictable. Data mining builds models to identify patterns among the variables in a data set. Some of these patterns are predictive (projecting future values), whereas others are explanatory (explaining the interrelationships among the variables) [7]. Using the created data set, we identify predictive and explanatory patterns. We predict the number of cases for subsequent years, and we study which socioeconomic indicators are more related to the registered hypertension cases in Mexico.

**Table 4.** Clustering algorithms precision evaluated by the the error range for k-means and by the accuracy given by correlation of distances for HCA

| | Precision | |
|---|---|---|
| **Year Data Set** | **k-means Error** | **HCA Accuracy** |
| 2008 | 11769.69 | 0.7854 |
| 2010 | 10436.07 | 0.7902 |
| 2012 | 9526.41 | 0.7658 |
| 2014 | 9991.22 | 0.7684 |
| 2016 | 8348.98 | 0.7925 |
| 2018 | 9780.10 | 0.7932 |

### 3.3.1 Forecast of the Number of Cases

Using the historical data since 2000, we predict the years 2019 and 2020 for each estate of Mexico and for the total number of detected cases in Mexico. We compare ARIMA [16] and CNN-QR [3] to forecast the number of hypertension cases.

Time series analysis can be viewed as the task of detecting patterns in data and predictability of values. Pattern detection can include: trend, cyclical, periodic, or outliers [9]. ARIMA obtains the representation of the series in terms of the temporal interrelation of its elements, characterizing the series as sums or differences, weighted or not, of random variables or the resulting series.

The fundamental element when analyzing the properties of a time series is the autocorrelation coefficient, which measures the degree of linear association that exists between observations separated by $k$ periods. The ARIMA model is commonly used in infectious disease time series prediction[23], and recently used for the prediction of COVID-19 cases [12].

Convolutional Neural Networks (CNN) are the evolution of Artificial Neural Networks, they use supervised learning that processes their layers to identify different characteristics in the inputs to identify objects. The CNN contains several specialized hidden layers with a hierarchy that become more specialized until reaching deeper layers.

Using causal CNN, the Convolutional Neural Network - Quantal Regression (CNN-QR) is an algorithm for predicting scalar time series. This supervised learning algorithm trains a collection of time series and uses a quantal decoder to make probabilistic predictions.

### 3.3.2 Correlation of BPH with Socioeconomic Factors

We used clustering for the purpose of analyzing the socioeconomic characteristics associated with the HBP cases detected. Clustering is used to group existing data whose common characteristics are unknown or are to be discovered.

The clusters generated are based on the similarity between states of the number of the proportion of people (since they are normalized data) that are part of these indicators. Two clustering algorithms were tested: k-means and hierarchical clustering.

*k-means* groups objects into $k$ groups depending on their characteristics. *k-means* clustering is an unsupervised learning technique to classify unlabeled data by grouping them by features, rather than pre-defined categories. The variable $k$ represents the number of groups or categories created.

The goal is to split the data into $k$ different clusters and report the location of the center of mass for each cluster. Then, a new data point can be assigned a cluster (class) based on the closed center of mass. The main advantage is that the machine creates its own clusters based upon empirical proofs, rather than assumptions.

Hierarchical Clustering (HCA) constructs a tree that represents the similar relationships between the different elements. The exploration of all possible trees is computationally intractable. We used agglomerative hierarchical clustering, it starts with as many clusters as there are individuals and consists of forming groups according to their similarity.

According to the selected IMSS data, there are 46 different socioeconomic indicators. Let $m$ be the number of socioeconomic indicators such that $m = 46$, and $S = \{s_1, s_2, ..., s_m\}$ be the set of socioeconomic indicators The best number of clusters $n$ was obtained by the elbow method. In this case $n = 5$, such that $C = \{C_1, C_2, C_3, C_4, C_5\}$. Each generated cluster $C_i$ has a vector of weights $W_{c_i} = \{w_{c_i s_1}, w_{c_i s_2}, ..., w_{c_i s_m}\}$, where each $w_{c_i s_j}$ represents the weight (impact) of indicator $s_j$ in the cluster $c_i$.

Let the cluster $c_x \in C$ have the highest weight associated with the number of BPH-identified cases. On the other side, let $w_{c_y s_z}$ be the highest weight of the socioeconomic indicator $s_z \in S$ in cluster $c_y \in C$. If $c_y$ is the same cluster as $c_x$, then it is established that the indicator $s_z$ is strongly related to the cases of hypertension registered that year.

**Table 5.** Recall of socioeconomic indicators in the same cluster (two cluster algorithms) compared with PCC

| | Recall | |
|---|---|---|
| **Year Data Set** | **k-means** | **HCA** |
| 2008 | **0.9166** | 0.8888 |
| 2010 | **0.9000** | 0.8750 |
| 2012 | 0.6000 | **0.8888** |
| 2014 | 0.9090 | **1.0000** |
| 2016 | 0.4444 | **1.0000** |
| 2018 | **0.8888** | 0.7500 |

## 4 Generated Data Set

We cover from the year 2000 to 2019, containing data related to Hypertension, nutrition, and socioeconomic indicators. Those repositories that contemplated less than three years of registration were eliminated since, based on the surveys, information was collected at intervals of two years. Finally, the analysis set was reduced to 46 databases and similar databases were integrated in 27 final CSV files.

The data was explored in search of cases such as repeated records, incomplete records, repeated fields, and fields with invalid data (null) or unintelligible content. Table 2 shows the numer of NaN and Nulls detected in the HAS datasets. To unify the information, the types of data used in different databases for fields with the information of the same type were identified, for example, for fields such as date were used, data types such as: $date$, $int$ or $string$.

To get data integrity, some values were modified; for example, in the databases of ENSANUT, there are interviews of patients, they used the value ”0” for negative responses in most of the cases but in 2016 and 2018 they used ”2”. Also, the periodicity of the records was evaluated to enter into the final system only those databases whose records coincided in the periods. For data integration, the datasets of the same institution were integrated as follows:

– From ENSANUT, the datasets of Surveys of Tennagers and Adults were integrated by years.

– From ENSANUT, the datases about Anthropometry Adults, Frequency of Adult Consumption, Food Safety, Information about the Household, and Information about the members of the household were joined considering a common number (folio) of register.

– From INEGI, we only consider the registers with direct cause of death of cardiac problems, the datasets of each year were integrated in one and a column of ”year of registration” was added.

– From the socioeconomic data from datos.gob, a manual exploration of 27 CSV files was performed, we integrated and merged the databases considereing to regional (municipality, state) aspect.

Then the data set includes the tables:

– IMSS with: year, state, and number of cases.

– INEGI-deaths with: registration entity, living entity, municipality, number of cases.

– INEGI with: registration entity, age, municipality of registration, detailed cause of death, mexican list of cause of death, sex, day of registration, month of registration, year of registration, occupation, escolarity, CIE detailed cause by chapter, CIE detailed cause by group, list of tabulation for CIE mortality, mexican list of disease.

– Socioeconomic-datos.gob with categories (with $n$ number of columns related to that category) of: percentage of number of people-average poverty deficiencies (16), Poverty intensity depth indicators (13), C4A percentage of number of people in need - average poverty by state (4), C4B percentage of number of people in need - average poverty by state (4), social deprivation indicators by state percentage (10), social deprivation indicators by state (10), components of social deprivation - Percentage of number of people (28), Measures of depth of poverty intensity by state A (7), Measures of depth of poverty intensity by state B (5), Total current income per capita according to income source A (40), Total current income per capita according to income source B , Percentage of number of people with average poverty deficiencies according to sex (17), Percentage of number of people with average poverty deficiencies under 18 years (17), Percentage of number of young people with average poverty deficiencies (17), Percentage of number of old people with average poverty deficiencies (17), Percentage of number of people with average poverty deficiencies according to etnic (17), Percentage of number of people with average poverty deficiencies with indigenous language (17), Percentage of number of people with average poverty deficiencies with disability (17), Percentage of number of people with average poverty deficiencies according to their residence (17), Percentage of number of people with average poverty deficiencies for state A (7), Percentage of number of people with average poverty deficiencies for state B (17), Percentage of number of people with average poverty deficiencies for state C (7), Percentage of number of people with average poverty deficiencies for state D (9), Percentage of number of people with average poverty deficiencies for state E (9), Percentage of number of people with average poverty deficiencies for state F (9), Press release tables with poverty information A (17), and Press release tables with poverty information B (3).

- ENSANUT-Nutrition with: entity, sex, age, weight, high, waist circumference, IMC, category of IMC, worry about food at home, home without food, homer without healthy food, low variety of food, one time food missing, eat less than you should, hungry, home with persons with less that 18 years, detected HAS, house construction material, house floor construction material, kitchen room, sleeping rooms, numer of rooms, electric light, water, number of days with eater, type of toilet, share bathroom, toilet flush type, cooking fuel, trash, owner, construct, were to cook, stove type, heating, number of bulb, more houses, car, truck, moto, other, TV, payed TV, radio, consols, iron, blender, refrigetaron, washing machine, computer, internet, microwave, phone, water tank, cistern, light, celphone, air conditioner, native, native language, school grade, civil status, number of persons at home, and share expenses of food.

Subsequently, the final representation of the data was generated in JSON, which is a lightweight data exchange format completely independent of language. The new representation allows hierarchical visualization and makes the data easier to read. This representation was chosen in order to speed up computational processes.

### 4.1 Forecast and Correlation with Socioeconomic Factors

A pattern means that the data are correlated, have a relationship, or are predictable. Data mining builds models to identify patterns among the variables in a data set. Some of these patterns are predictive (projecting future values), whereas others are explanatory (explaining the interrelationships among the variables) [7].

Using the created data set, we identify predictive and explanatory patterns. We predict the number of cases for subsequent years, and we study which socioeconomic indicators are more related to the registered hypertension cases in Mexico.

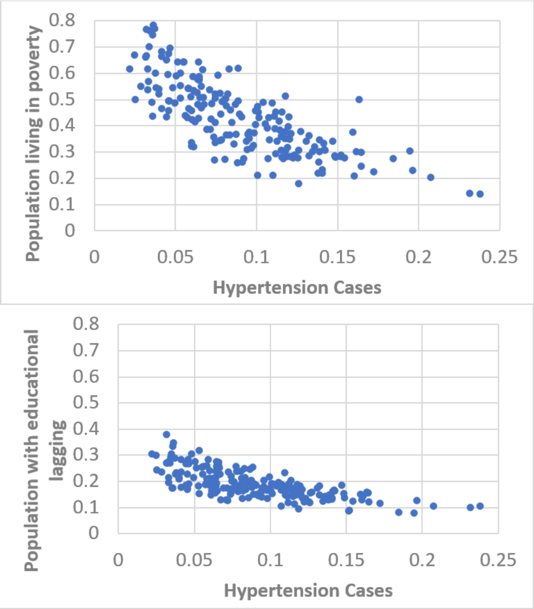

**Fig. 2.** Correlation of the number of BPH cases and the indicator by state and year

## 5 Evaluation of Patterns in Hypertension Data

There are two main patterns presented in this paper. The first one is focused on the forecast of cases of hypertension with two models that are evaluated with RMSE and WAPE. The second one evaluates the correlation of BPH with socioeconomic indicators using clustering (k-means and HCA) and identifying the indicators with the highest weight for the cluster with more HAS; we also used Pearson Correlation Coefficient to identify the principal indicators in order to compare them with the results given by the clusters using Recall.

### 5.1 Metrics

Two analysis are presented. To evaluate the prediction of cases we used RMSE and WAPE. To evaluate the correlation of BPH with socioeconomic factors we use recall and PCC.

#### 5.1.1 Root Mean Square Error (RMSE)

It is a standard way to measure the error of a model in predicting quantitative data. It compares the observed value $y_i$ and the forecast value $\hat{y}_i$.

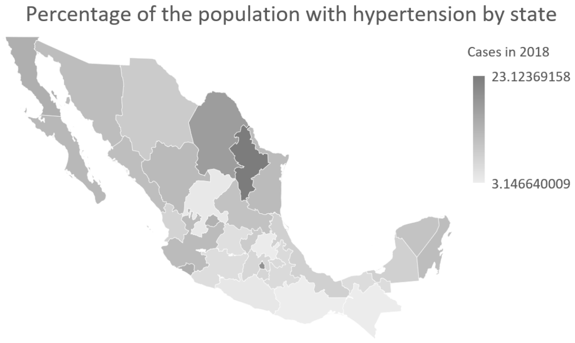

**Fig. 3.** Percentage of population with BPH for each state. Darker states have more hypertension

This metric uses the squared value of residuals, amplifying the impact of outliers, then:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}(y_i - \hat{y_i})^2}. \tag{1}$$

Such that $i = 1, ..., N$, where $N$ is the number of data point. The smaller the RMSE value, the better the predictive accuracy of the model.

### 5.1.2 WAPE o Weighted Absolute Percentage Error

It is one of the most common methods to measure the overall deviation of forecast values from observed values. It measures the percentage error of the sum of the observed values versus the sum of predicted values. A lower value indicates a more accurate model. Let $y_{i,t}$ be the observed value at point $(i, t)$ and $\hat{y_{i,t}}$ the predicted value at point $(i, t)$:

$$WAPE = \frac{\sum_{i,t}|y_{i,t} - \hat{y_{i,t}}|}{\sum_{i,t}|y_{i,t}|}. \tag{2}$$

### 5.1.3 Recall

The completeness metric will inform us of the number of samples that the machine learning model can identify. For this paper, recall refers to the percentage of important indicators we identify:

$$recall = \frac{TP}{TP + FN}, \tag{3}$$

where, $TP$ refers to True Positive where the predicted elements are the real ones, and $FN$ is the False Negative, which are the real elements that we did not predict.

### 5.1.4 Pearson Correlation Coefficient (PCC)

It is a measure of linear correlation between two sets of data. By definition PCC $\rho$ is the covariance of the two variables $(X, Y)$ divided by the product of their standard deviations:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, \tag{4}$$

where, $\sigma_X$ is the standard deviation of X, $\sigma_Y$ is the standard deviation of Y, $cov$ is the covariance given by:

$$cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)], \tag{5}$$

where $\mu_X$ is the mean of $X$, $\mu_Y$ is the mean of $Y$, and $\mathbb{E}$ is the expectation.

## 5.2 Evaluation of the Prediction of the Number of Cases of BPH

To develop the prediction of the number of HBP cases, several tests were performed by generating the models, as it is known efficiency decreased as the prediction range increased (backtest window). The predictions made at one year showed the lowest WAPE and RMSE, while for predictions of 2, 3, and 5 years they increased.

Table 1 shows this behavior and its error parameters. WAPE is more robust to outliers than Root Mean Square Error (RMSE) because it uses the absolute error instead of the squared error.

For the evaluation, a one-year backtest window was performed, i.e., the algorithm predicted 2019, which is the year for which data is available.

It should be noted that the same tests were performed for the prediction of 2020; however, since it was a pandemic year, the cases of hypertension recorded were below those recorded 20 years ago; since it was an atypical year, the data obtained from official sources were not consistent.

We compare two algorithms: Mean ARIMA and CNN-QR (Convolutional Neural Network - Quantile Regression). The results given by ARIMA were appropriate compared with some others; nevertheless, CNN-QR showed better results, as we can see in Table 3.

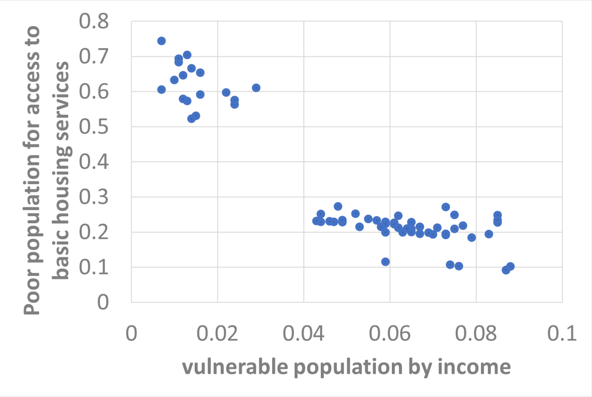

**Fig. 4.** Clusters given by the correlation between

### 5.3 Evaluation of the Correlation of BPH with Socioeconomic Indicators

In order to compare the number of HBP cases with socioeconomic indicators, the required data sets were normalized. We used clustering to identify the cluster where hypertension is more important (highest weight). Then, we identify the socioeconomic indicators with the highest presence in the same cluster, such that those indicators are associated with hypertension as it was explained in Section 3.3.2.

We evaluated two clustering methods: Hierarchical Clustering (HCA) and k-means. Let $m$ be the number of variables $v$, for each observation $e$ there is an input vector such that $e_i = \{v_1, v_2, ..., v_m\}$, and there is an observation per year and state in Mexico.

The input vector for each observation has 47 variables, such that $m = 47$ : 46 socioeconomic characteristics and one that corresponds to the percentage of HAS cases per state. To balance the weights, the HAS cases component has a weight of 46, while the socioeconomic characteristics were assigned to one unit of weight, thus balancing the HAS with the 46 indicators.

We compare the indicators $S$ of both algorithms and validate them with the Pearson correlation coefficient, which measures the statistical relationship, or association, between two continuous variables. It returns a value of between $-1$ and $+1$, where +1 represents the highest positive correlation.

For the HCA the distances between each observation in the hierarchical grouping, defined by the linkage matrix, were calculated and the correlation of distances were obtained to determine the accuracy. While for K-means, the sum of squared distances from the center to the nearest cluster were calculated, which can be interpreted as the error range of the same model. The accuracy of hierarchical clustering is measured through the generated distance matrix and the original data, so the variation in the number of clusters did not greatly affect the result. Within these tests, it was decided to use five clusters for both groupings.

The expected value for hierarchical clustering accuracy is 1, while for K-means error it is zero (it is important to remember that the latter is a sum of squared distances). The accuracy obtained was above 0.75 (75% success rate), i.e., for the 32 data vectors entered corresponding to each state, at least 24 were correctly clustered according to their socioeconomic characteristics. The classification results are presented in Table 4.

Once the clusters are obtained, we evaluated the output vectors $w$ for each $k$ cluster, such that each $w_{ij}$ represents the weight given to a each variable $v_j \in V$ for each cluster $C_i$ where $1 \leq i \leq k$ and $1 \leq j \leq m$ for $m$ number of variables. Then, each socioeconomic indicator $v_j$ has a weight associated to each cluster such that, $V_{v_j} = \{w_{1v_j}, w_{2v_j}, ..., w_{kv_j}\}$.

Let $C_c$ be the cluster with the highest presence of BPH, if $w_{c_c v_j}$ is the highest weight in $V_{v_j}$ then $v_j$ is an important indicator for $C_c$. In order to evaluate it, we compare the important indicators to the top five important indicators given by PCC. The recall of this comparison is in Table 5. For both algorithms the recall is high because it is near to 1.

Some of the most frequent socioeconomic indicators correlated with BPH are population vulnerable by income, non-poor and non-vulnerable population, contribution of educational backwardness to poverty, contribution of access to health care to poverty, contribution of access to social security to poverty, contribution of access to food to poverty, population affiliated to IMSS, population with private medical insurance, population with access to indirect social security health services and population food insecurity.

## 6 Discussion and Further Analysis

Data were collected, integrated, normalized, and preprocessed by the team. This resulted in an unpublished dataset, so the results of socioeconomic factors correlated with hypertension were validated using a ranking derived from the Pearson coefficient through recall. There are some other socioeconomic indicators that are not predominant in the cluster with the highest registered hypertension, like the population in poverty, with educational lagging, and without access to social security. It was deduced that hypertension itself could be associated with these factors; however, the databases considered are those of **registered cases** of hypertension.

Therefore, in people with the aforementioned indicators, the disease has not yet been identified, but that does not mean that they do not have it. An early diagnosis of this type of disease is vital; people with deficiencies often do not have access to medical services and therefore are not detected in time.

The final step of the data mining process is the data representation. Some graphic patterns were generated. Figure 2 shows the correlation of the number of BPH cases and two indicators. Each point represents a state and a year located on the plane that correlates the number of hypertension cases and the indicator.

Once we are evaluating the number of cases of BPH in Mexico for each state, Figure 3 shows the country; darker states have the highest register of BPH cases, like Mexico City in the middle of the map and Nuevo León in the northeast.

The generated data set could also be used to analyze some correlations inside the socioeconomic factors; for example, Figure 4 shows two evident clusters if we correlate vulnerable population by income and poor population for access to basic housing services.

## 7 Conclusions

The initial phase of this project facilitated the creation of a comprehensive database consolidating data from diverse public sources on High Blood Pressure (HBP or hypertension), including socioeconomic indicators of the patients. The collected data underwent a meticulous process of organization and standardization, resulting in the establishment of a robust data repository.

The standardized data enabled the prediction of registered cases of high blood pressure and an assessment of the potential influence of factors not traditionally considered by medical professionals in the development of this disease. The algorithm takes into account variables such as food insecurity, susceptibility due to income inadequacy, and affiliation to health services.

Drawing insights from information spanning the past 10 years, a projection was generated concerning the trajectory of HTN diagnoses, exhibiting promising outcomes until 2019. However, the results for 2020 were compromised due to the lack of records during the COVID-19 pandemic.

## Acknowledgments

We appreciate the hard work of the Telematics Engineering Students Erika De-la-Fuente-Castañeda and Leonardo Aguilar-Rosas from the Instituto Politécnico Nacional (UPIITA-IPN), Mexico City, Mexico. This work was supported by the Proyecto Individual 20231347 and by the Proyecto Multidisciplinario 20231738 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional.

## References

1. **Agarwal, M., Agrawal, S., Garg, L., Lavie, C. J. (2017).** Relation between obesity and survival in patients hospitalized for pulmonary arterial hypertension (from a nationwide inpatient sample database 2003 to 2011). The American Journal of Cardiology, Vol. 120, No. 3, pp. 489–493. DOI: 10.1016/j.amjcard.2017.04.051.

2. **Alfonso-González, W. (2013).** Predicción de la evolución hacia la hipertensión arterial en la adultez desde la adolescencia utilizando técnicas de aprendizaje automatizado. Trabajo de Diploma. Facultad de Matemática, Física y Computación, Universidad Central Marta Abreu De las Villas.

3. **AWS Service Team (2020).** Time series forecasting principles with Amazon forecast. Amazon Web Service.

4. **Boo, S., Yoon, Y. J., Oh, H. (2018).** Evaluating the prevalence, awareness, and control of hypertension, diabetes, and dyslipidemia in Korea using the NHIS-NSC database: a cross-sectional analysis. Medicine, Vol. 97, No. 51, pp. e13713. DOI: 10.1097/md.0000000000013713.

5. **Campos, I., Hernández, L., Rojas, R., Pedroza, A., Medina, C., Barquera, S. (2013).** Hipertensión arterial: Prevalencia, diagnóstico oportuno, control y tendencia en adultos mexicanos. Salud Pública de México, Vol. 55, No. 2.

6. **Datos Abiertos (2023).** Descubre datos abiertos de tu gobierno. datos.gob.mx.

7. **Delen, D. (2020).** Predictive analytics: Data mining, machine learning and data science for practitioners. FT Press.

8. **Dritsas, E., Fazakis, N., Kocsis, O., Fakotakis, N., Moustakas, K. (2021).** Long-term hypertension risk prediction with ml techniques in elsa database. Learning and Intelligent Optimization: 15th

International Conference, pp. 113–120. DOI: 10.1007/978-3-030-92121-7_9.

9. **González-Castellanos, M., Soto-Valero, M. (2013).** Minería de datos para series temporales. Facultad de Matemática, Física y Computación, U.C Marta Abreu De Las Villas.

10. **Grotto, I., Huerta, M., Sharabi, Y. (2008).** Hypertension and socioeconomic status. Current Opinion in Cardiology, Vol. 23, No. 4, pp. 335–339. DOI: 10.1097/hco.0b013e3283021c70.

11. **INEGI (2023).** Instituto Nacional de Estadística y Geografía. www.inegi.org.mx.

12. **Khan, F. M., Gupta, R. (2020).** ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. Journal of Safety Science and Resilience, Vol. 1, No. 1, pp. 12–18. DOI: 10.1016/j.jnlssr.2020.06.007.

13. **Kim, S., Kim, H. (2016).** A new metric of absolute percentage error for intermittent demand forecasts. International Journal of Forecasting, Vol. 32, No. 3, pp. 669–679. DOI: 10.1016/j.ijforecast.2015.12.003.

14. **Kopeć, G., Kurzyna, M., Mroczek, E., Chrzanowski, Ł., Mularek-Kubzdela, T., Skoczylas, I., Torbicki, A. (2019).** Database of pulmonary hypertension in the polish population (BNP-PL): Design of the registry. Kardiologia Polska (Polish Heart Journal), Vol. 77, No. 10, pp. 972–974.

15. **Leng, B., Jin, Y., Li, G., Chen, L., Jin, N. (2015).** Socioeconomic status and hypertension: A meta-analysis. Journal of hypertension, Vol. 33, No. 2, pp. 221-229. DOI: 10.1097/HJH.0000000000000428.

16. **Mardianto, I., Muhamad-Ichsan, G., Dedy, S., Abdul, R. (2020).** Comparison of rice price forecasting using the arima method on amazon forecast and sagemaker. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), Vol. 4, No. 3, pp. 537–543. DOI: 10.29207/resti.v4i3.1902.

17. **Pickering, T. (1999).** Cardiovascular pathways: Socioeconomic status and stress effects on hypertension and cardiovascular function. Annals of the New York Academy of Sciences, Vol. 896, No. 1, pp. 262–277. DOI: 10.1111/j.1749-6632.1999.tb08121.x.

18. **Pérez-Fernández, G. A., Grau-Abalo, C. R. (2012).** Predicción de la evolución hacia la hipertensión arterial en la adultez desde la adolescencia. Revista Cubana de Informática Médica, Vol. 4, No. 1, pp. 43–53.

19. **Ramezankhani, A., Kabir, A., Pournik, O., Azizi, F., Hadaegh, F. (2016).** Classification-based data mining for identification of risk patterns associated with hypertension in middle eastern population: A 12-year longitudinal study. Medicine, Vol. 95, No. 35, pp. e4143. DOI: 10.1097/md.0000000000004143.

20. **Saleh, B. J., Saedi, A., al-Aqbi, A., Abdalhasan-Salman, L. (2020).** Analysis of weka data mining techniques for heart disease prediction system. International journal of medical reviews, Vol. 7, No. 1, pp. 15–24.

21. **Secretaría de Salud, Enfermedades No Transmisibles Situación y Propuesta de Acción (2018).** Una perspectiva desde la experiencia de México, 2018. Primera edición. México: Secretaría de Salud.

22. **Secretaría de Salud (2019).** 17 de mayo día mundial de la Hipertensión Arterial. Content/uploads/2019/05/Hipertensio_2019.pdf.

23. **Wang, Y., Shen, Z., Jiang, Y. (2018).** Comparison of ARIMA and GM(1,1) models for prediction of hepatitis B in China. PLOS ONE, Vol. 13, No. 9, pp. e0201987. DOI: 10.1371/journal.pone.0201987.

24. **Xu, X., Kawakami, J., Millagaha-Gedara, N. I., Riviere, J. E., Meyer, E., Wyckoff, G. J., Jaberi-Douraki, M. (2021).** Data mining methodology for response to hypertension symptomology—application to COVID-19-related pharmacovigilance. eLife, Vol. 10, pp. e70734. DOI: 10.7554/elife.70734.

*Article received on 10/12/2023; accepted on 21/10/2024.*
**Corresponding author is Bella Martinez-Seis.*