2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Hypertension and its Relationship with Socioeconomic Factors in Mexico Using Clustering Techniques

1st Obed Casillas-Baltazar
UPIITA-IPN
Intituto Politécnico Nacional
Mexico City, Mexico
ocasillas1800@alumno.ipn.mx

2nd Obdulia Pichardo-Lagunas
UPIITA-IPN
Intituto Politécnico Nacional
Mexico City, Mexico
opichardola@ipn.mx

3rd Bella Martinez-Seis
UPIITA-IPN
Intituto Politécnico Nacional
Mexico City, Mexico
bcmartinez@ipn.mx

***Abstract*—According to Ministry of Health in Mexico, Hypertension, commonly referred to as High Blood Pressure (HBP), continues to rank among the foremost ten causes of mortality in Mexico. This document describes the methodology for unearthing correlations between non-clinical variables and HBP, utilizing data clustering techniques in a data set derived from diverse Mexican institutions.**

***Index Terms*—Hypertension,Socioeconomic Factors, Clustering Techniques**

I. Introduction

Over the past twenty years, the global incidence of Non-Communicable Diseases (NCDs) has been on the rise. This issue is particularly pronounced in low- and middle-income countries, such as Mexico, where 75% of the deaths are associated with NCDs including Systemic Arterial Hypertension (SAH) [2]. Notably, systemic arterial hypertension has demonstrated a notable epidemiological shift in recent times. It is thought that together with the congenital factor of the disease, there are environmental and socioeconomic variables that can affect its development in the population.

II. Patterns of Associations within Hypertension and Socioeconomic Data

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) [1]. We integrated a data set from health institutions; then we study which socioeconomic indicators are more related to the registered hypertension cases in Mexico.

First, we will describe the data collection, selection, and integration of data related to hypertension and socioeconomic variables from health institutions in Mexico. Then, we propose a methodology using clustering in order to find correlations between Hypertension and Socioeconomic Data. Finally, we evaluate them.

A. Data collection and selection for HBP and Prepossessing

The project encompasses two primary data categories: hypertension and socioeconomic variables that were collected from Mexican institutions. Patient records for hypertension are upheld by the Mexican government and health institutions such as IMSS, ISSSTE, and ENSANUT. Socioeconomic data originates from INEGI and the Mexican government's datos.gob open data platform.

Over 175 sources from ENSANUT, IMSS, ISSSTE, and the Health Sector were collected. We got 100,000+ records with 1,000+ attributes encompassing 2000 to 2020. Data sets include information on hypertension and socioeconomic indicators. A comprehensive exploration identified recurring entries, incomplete data, duplicated fields, and instances of invalid or unclear content. Data types were standardized. Data periodicity was evaluated, leading to the inclusion of databases with overlapping record periods into the final system. Finally, we got a unified data set in JSON format.

B. Clustering for correlation

Let $S = \{s_1, s_2, ..., s_m\}$ be the set of socioeconomic indicators; according to the selected data there are 46 different indicators, such that $m = 46$. 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. Among them, the cluster c_x, such that $0 \leq x \leq N$, has the highest weight associated with the number of BPH-identified cases. On the other hand, the weights of each socioeconomic indicator s_j for each cluster c_i are evaluated. We obtain the cluster c_y, such that $0 \leq y \leq N$, in which the socioeconomic indicator s_j is dominant. If c_y is the same cluster as c_x, then it is established that the indicator s_j is strongly related to the cases of hypertension registered that year.

Identify applicable funding agency here. If none, delete this.

978-0-7381-4408-5/23/$31.00 ©2023 IEEE

Two clustering algorithms were tested: k-means and hierarchical clustering. We used *k-means*,it groups objects into k groups depending on their characteristics. Hierarchical Clustering constructs a tree that represents the similar relationships between the different elements. We compare the indicators j 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.

III. Evaluation of the Correlation of BPH with Socioeconomic Indicators

We evaluate 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.

In order to compare the number of HBP cases with socioeconomic indicators, the required data sets were normalized. We test it for each of the 32 states in Mexico. We evaluated two clustering methods: Hierarchical Clustering (HCA) and k-means.

For k-means, the sum of squared distances from the center to the nearest cluster was calculated, which can be interpreted as the error range of the same model. For the HCA the distances between each observation in the hierarchical grouping, defined by the linkage matrix, were calculated; and the correlation of distances was obtained to determine the accuracy. The accuracy of HCA is measured through the generated distance matrix and the original data. 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 0.7825 (78.25% success rate), i.e., for the 32 data vectors entered corresponding to each state, at least 25 were correctly clustered according to their socioeconomic characteristics. The classification results are presented in Table I.

TABLE I
Clustering algorithms precision evaluated by the the error range for k-means and by the accuracy given by correlation of distances for HCA

	Precision	
	k-means	HCA
Year Data Set	Error	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

Once the clusters are obtained, we evaluated the output vectors W for each c_i cluster and for each socioeconomic indicator s_j. 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 II. For both algorithms, the recall is high because it is near 1.

TABLE II
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

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.

IV. CONCLUSIONS

The initial phase of this project involved establishing a unified database by amalgamating data from various publicly accessible sources concerning High Blood Pressure (HBP or hypertension) and socioeconomic indicators. Leveraging this standardized data enabled the prediction of registered hypertension cases and an assessment of the potential impact of non-medical factors on its development. Variables such as food insecurity, income vulnerability, and healthcare affiliation were examined for their potential influence.

References

[1] Delen, Dursun. Predictive Analytics: Data Mining, Machine Learning and Data Science for Practitioners. FT Press, 2020.

[2] Secretaría de Salud, Enfermedades No Transmisibles Situación y Propuesta de Acción: Una Perspectiva desde la Experiencia de México, 2018. Primera edición. México: Secretaría de Salud, 2018.