Article Journal-Mathematical and Quantitative Methods
June, 2019 Vol.3 No.4 1-8

Implementation of the ID3 algorithm for the generation of a decision tree with food health data from the State of Guerrero, Mexico

Implementación del algoritmo ID3 para la generación de un árbol de decisión con datos de salud alimenticia del Estado de Guerrero, México

GALLARDO-BERNAL, Iván†*, MOLINA-ÁNGEL, Félix, HERNÁNDEZ-HERNÁNDEZ, José Luis and HERRERA-MIRANDA, Israel

Universidad Autónoma de Guerrero. Escuela Superior de Gobierno y Gestión Pública.

ID 1st Author: *Iván, Gallardo-Bernal* / **ORC ID**: 0000-0002-1596-6786, **CVU CONACYT ID**: 613169

ID 1st Coauthor: *Félix, Molina-Ángel* / **ORC ID**: 0000-0002-7834-1812, **CVU CONACYT ID**: 478151

ID 2nd Coauthor: *José Luis, Hernández-Hernández* / **ORC ID**: 0000-0003-0231-2019, **CVU CONACYT ID**: 294222

ID 3rd Coauthor: *Israel, Herrera-Miranda* / **ORC ID**: 0000-0001-8031-797X, **CVU CONACYT ID**: 299348

DOI: 10.35429/JMQM.2019.4.3.1.8 Received March 28, 2019; Accepted June 30, 2019

Abstract

Mexico faces a significant public health problem related to the eating habits of the population. More than 50% of the world's obese people live in 10 countries, including Mexico (Ng et al., 2014). In 2014, Mexico had the highest obesity prevalence rate in Latin America (PAHO, 2016a). The prevalence of obesity in adults in the state of Guerrero in 2012 was 71.5% (Barquera, 2012) The objective of this research was to generate a diagnosis of these conditions in both urban and rural locations in the seven regions of the state of Guerrero. The information was analyzed using an information system called SIOB, which is a technological tool developed to collect information from the population with a sample of 23,000 people, users of the public health sector of the state of Guerrero. We used the Weka tool of the University of Waikato (Eibe et al., 201d6) to generate a visible decision tree that shows the prevalence of these conditions in the state of Guerrero. It is considered that these visualization tools can help in the implementation of public policies that contribute to the prevention of these chronic degenerative diseases.

ID3, Decision Tree, Obesity

Resumen

México enfrenta un importante problema de salud pública, relacionado con los hábitos alimenticios de la población. Más del 50 % de las personas obesas del mundo viven en 10 países, incluido México (Ng et al.,2014). En 2014, México tenía la tasa de prevalencia de obesidad más alta de América Latina, (PAHO, 2016a). La prevalencia de la obesidad en adultos en el estado de Guerrero en 2012 fue de 71.5% (Barquera, 2012). La presente investigación tuvo como objetivo generar un diagnóstico de estos padecimientos tanto en las localidades urbanas como rurales de las siete regiones del estado de Guerrero. La información se analizó utilizando un sistema de información denominado SIOB, que es una herramienta tecnológica desarrollada para recabar información de la población con una muestra de 23,000 personas, usuarias del sector público de salud del estado de Guerrero. Se utilizó la herramienta Weka de la Universidad de Waikato (Eibe, et al., 2016) para generar un árbol de decisión visualizable que muestra la prevalencia de estos padecimientos en el estado de Guerrero. Se considera que estas herramientas de visualización pueden auxiliar en la instrumentación de políticas públicas que contribuyan a la prevención de estas enfermedades crónico-degenerativas.

ID3, Árbol de Decisión, Obesidad

Citation: GALLARDO-BERNAL, Iván, MOLINA-ÁNGEL, Félix, HERNÁNDEZ-HERNÁNDEZ, José Luis and HERRERA-MIRANDA, Israel. Implementation of the ID3 algorithm for the generation of a decision tree with food health data from the State of Guerrero, Mexico. Journal-Mathematical and Quantitative Methods. 2019. 3-4: 1-8

* Correspondence to Author (email: drivangallardo@gmail.com)
† Researcher contributing first author

Introduction

The role of technology in the Health Sector Plan 2013-2018 in the ENSANUT MC (2016) maintains that health is an indispensable condition for people's well-being and is one of the fundamental components of human capital (Correa, 2008). Likewise, INEGI (2009) and CONAPO (2013) maintain that advances in the population's health status are mainly due to the conditions in which people are born, grow, live, work and age. Progress in education, income level, food, housing conditions, among others, are conditions that influence people's health status and are known as social determinants of health.

Concerning progress in the diet of the population in Mexico from the nutritional point of view, the Sectoral Health Plan (PROSESA, 2013-2018) reports the following: That one of the leading chronic diseases that cause a significant number of deaths in both women and men from age 20, exacerbated in those over 65, is the prevalence of overweight and obesity, as well as unhealthy lifestyles, causing the leading causes of death among the adult population such as diabetes mellitus and ischemic heart disease.

Besides, the Ministry of Health, through a National Survey on Health and Nutrition in the Midwest (ENSANUT, 2012), conducted a study of the nutritional situation in four regions (North, Center, Mexico City, and South), their urban and rural locations.

This study shows that almost 4 out of 10 adolescents are overweight or obese, and seven out of 10 adults are overweight, with a combined prevalence of 72.5%, compared to 71.2% in 2012. On the other hand, Mexico belongs to the nations with the highest adult obesity in the world, this according to the Organization for Economic Cooperation and Development (OECD, 2017) and the countries with the highest prevalence are The United States, with 38.2%, Mexico, with 32.4% and New Zealand, with 30.7%. For this, it is projected that obesity in Mexico will increase to 39% by 2030.As a consequence, obesity is the leading risk factor for the development of type 2 diabetes, presenting a critical dimension, occupying the first place in the world, in terms of people who suffer from it within the age range of 20-79 years.

Mexico is characterized by notable differences in overweight or obesity between regions, urban/rural localities, and in the different socioeconomic levels and the states of very high marginalization such as Oaxaca, Chiapas, and Guerrero (Larose, 2005).

From the analysis carried out in the studies mentioned above, we knew the magnitude and distribution of the problem in our country, its association with other risk factors, and even some of its consequences stratified by region, socioeconomic level, and locality. From this knowledge of the situation that Mexico presents in terms of overweight and obesity, it has been possible to identify the need to have a diagnosis that contributes to pinpoint the profile of the health conditions of indigenous and non-indigenous populations that live in localities of high and very-high marginalization in Guerrero.

According to the World Health Organization (WHO) (2008), Guerrero is a state that continues to rank first in terms of obesity and overweight. This diagnosis was carried out with the goal of publicizing the state of nutrition safety in various entities in Mexico and thus be able to develop strategies for the prevention and control of this condition, as well as generate a tool to support decision-making that allows for the development of objective, effective, and efficient public policies that contribute to the eradication of the problem.

Implementation of the ID3 algorithm: Description of the Method

The first step in collecting the data was the creation of a system called SIOB for the collection, storage, and processing of information in the general hospitals and centers of the public health system in the state of Guerrero. It is worth mentioning that the state of Guerrero has approximately 1165 medical units. The SIOB system thus has information on the nutritional status of the population and identifies in which region of the state there is a higher concentration of overweight.

In Table 1, we show some data that the SIOB system collects for the seven regions of the state of Guerrero. This system was developed according to the norms of the World Health Organization (WHO) to calculate the Body Mass Index (BMI) and classifies the results of the sample by age and gender.

This standard suggests obtaining the BMI according to the age and gender of each patient (Gallardo 2015).

Attribute	Stored on SIOB Platform	Discretization for Arff
Gender	Masculine, Femenine	M, F.
Region	1,2,3,4,5,6,7	Acapulco, Center, Mountain, Tierra Caliente, North Zone, Costa Chica, Costa Grande
Diagnosis	Low Weight Mild malnutrition Moderate malnutrition Severe Malnutrition Normal Overweight Obesity Obesity I Morbidly Obese	– BP – DL – DM – DS – NM – OBD – OBU – OM – SP

Table 1 Information used for this study

The information generated by SIOB is produced from interviews with patients who attend the medical units of the State Health Services in the state of Guerrero. The diagnoses obtained by the system can be the following: Underweight (LP), Mild Malnutrition (ML), Moderate Malnutrition (MD), Severe Malnutrition (SD), Obesity (OBD), Normal (NM) First Grade Obesity (OBU) Morbid Obesity (OM), Overweight (SP).

The proposed sample for this diagnosis was the following:

- Population universe: 3, 533,251 People
- Geography: State of Guerrero
- Sample collected: 23,000 people
- Sample Valid for study 17,143
- Civilian population living in one of the 7 regions of the state of Guerrero.
- Instrument: Personal interview in medical units (SIOB)

Data Selection and Discretization

- 17,143 patients were selected, derived from the careful analysis, these data ensure the integrity of the information. (complete, without duplication, etc.)
- The selected patients are over 1.40 cm tall (this is used as a filter that helps us stratify the appropriate age segment for the study)
- All selected patients are over 19 years of age since, according to the WHO, the body mass index (BMI) has a generic way of being calculated.
- All patients are residents of one of the 7 regions of the state of Guerrero.
- All the data was captured from interviews with patients who attended the health sector consultation service.

This information was discretized in a flat text file so that it could be analyzed with the ID3 algorithm (Quinlan, 1996) and consequently generate a decision tree diagram that allowed us to visualize the processed data. The ID3 algorithm is used as a metric to select the best attribute to divide the data into homogeneous classes, which "learns" from the difference between the data to be analyzed. That is, a divide and conquer procedure, which maximizes the information obtained (Gallardo, 2015). This algorithm is used within the field of artificial intelligence to achieve a search for hypotheses or rules in it, given a set of examples issued by a set of data. The ID3 searches for the best attribute that is established by entropy, which measures the degree of organization of the system, choosing the attribute that provides a better gain of information.

Implementation of the ID3 algorithm

Data discretization and processing in the Waikato Environment for Knowledge Analysis (WEIKA) tool.

WEKA: Waikato Environment for Knowledge Analysis (Witten, I. H., Y Frank, 2005) is a tool that supports different standard data mining tasks, especially, data pre-processing, clustering, classification, regression, visualization, and selection. All Weka techniques are based on the assumption that the data are available in a flat file, where each data record is described by a fixed number of attributes, usually numerical or nominal, although other types are also supported.

ISSN 2531-2979

One of WEKA's requirements for optimal operation is that once discrete information has been achieved, the data must be transferred to a plain text file with the extension Arff. We call this procedure pre-processing (see Table 2).

Preparation (Arff) Attribute Relationship File Format.

```
%2. Sources:
%   (a) Iván Gallardo Bernal
%   (c) Date: Noviembre, 2018
%3. Past Usage:
%4. Relevant Information Paragraph:
%   paciente        Muestra
%   Género          Masculino, Femenino
%   Diagnóstico     Tipo de diagnóstico
%   Estatura        Estatura del paciente
%   Edad            Edad del paciente
% 5. Number of Instances: 17144
%   (instances completely cover the attribute space)
% 6. Number of Attributes: 2
% 7. Attribute Values:
%   Genero          1,0
%   Diagnóstico     1,0
%   Estatura        1,0
% 8. Missing Attribute Values: none
% Information about the dataset
% CLASSTYPE: nominal
% CLASSINDEX: last

@relation PACIENTE

@attribute Genero {1,0}
@attribute Diagnostico {1,0}
@attribute Estatura {1,0}
@attribute Edad {1,0.5,-1}

@data

0,1,1,-1
0,1,1,-1
0,1,1,-1
0,1,1,-1
0,1,1,-1
0,1,1,-1
0,1,1,-1
0,1,1,0,5
0,1,1,-1
0,0,1,0,5
1,0,1,-1
0,0,0,-1
0,0,1,1
0,0,1,0,5
0,0,0,-1
1,0,0,-1
1,0,0,-1
1,0,1,-1
```

Table 2 Plain Text Arff, Source: Prepared by the authors

Figure 1 shows the data entry interface of the Weka tool.

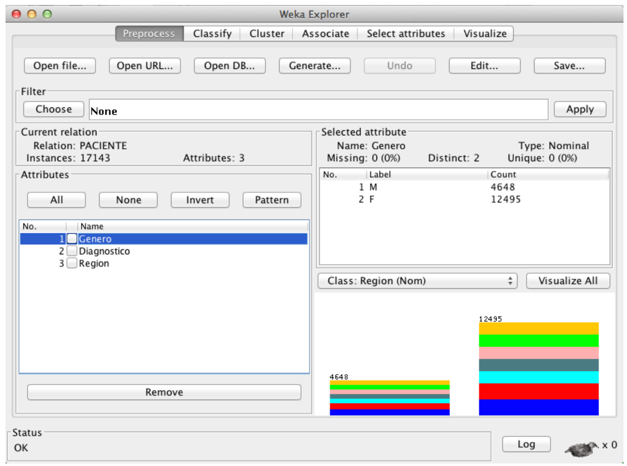

Figure 1 Data entry to the Weka tool
Source: (Weka, 2005)

Figure 1 shows the capture of the .Arff file for the program to identify the relationships of the attributes which will be processed with the ID3 algorithm.

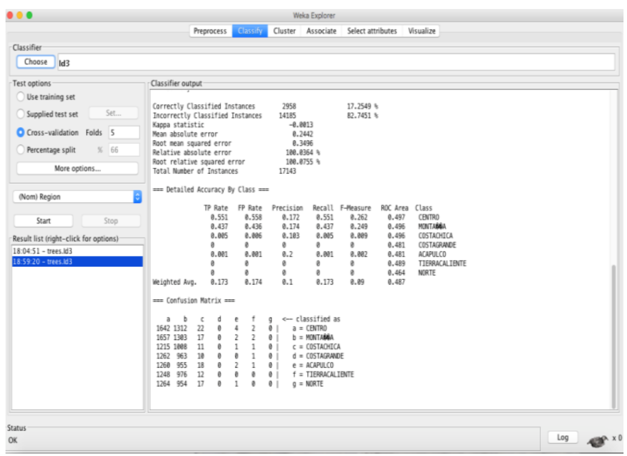

Figure 2 Data Processing in Weka
Source: Prepared by the authors from the data of the author

According to the information processing (Figure 2), the following data matrix was obtained and used as classifiers for the elaboration of the decision tree. (Table 3)

<table>
<tr><th>Gender = M</th><th>Gender = F</th></tr>
<tr><td>| DX = BP: CENTRO</td><td>| DX = BP:MONTAÑA</td></tr>
<tr><td>| DX= DL: CENTRO</td><td>| DX = DL: MONTAÑA</td></tr>
<tr><td>| DX = DM: MONTAÑA</td><td>| DX= DM: CENTRO</td></tr>
<tr><td>| DX = DS: CENTRO</td><td>| DX = DS: MONTAÑA</td></tr>
<tr><td>| DX = NR: MONTAÑA</td><td>| DX = NR: CENTRO</td></tr>
<tr><td>| DX = OBD: CENTRO</td><td>| DX = OBD: MONTAÑA</td></tr>
<tr><td>| DX = OBU: MONTAÑA</td><td>| DX = OBU: CENTRO</td></tr>
<tr><td>| DX = OM: CENTRO</td><td>| DX = OM: MONTAÑA</td></tr>
<tr><td>| DX = SP: MONTAÑA</td><td>| DX = SP: CENTRO</td></tr>
</table>

Table 3 Gender Confusion Data Matrix
Source: Prepared from the data of the authors

June, 2019 Vol.3 No.4 1-8

A data jumble matrix is a tool that allows the visualization of the performance of an algorithm used in supervised learning. One of the benefits of confounding matrices is that they make it easier to see if the system is confusing two classes. This tool is used within the scope of artificial intelligence, and its use is encompassed in the search for hypotheses or rules, given a set of examples. Table 4 shows the data and regions that will be processed for the creation of the decision tree.

Amount of data analyzed by region							Regions
a	b	c	d	e	f	g	Classification
1818	114	0	9	0	4	1	a: Center
1791	117	4	11	0	3	2	b: Mountain
1327	90	3	5	0	0	1	c: Costa Chica
1381	85	1	3	0	0	1	d: Consta Grande
1378	84	9	9	0	0	0	e: Acapulco
1365	86	6	2	0	3	0	f: Tierra Caliente
1374	85	4	7	0	1	0	g: North

Table 4 Confusion Matrix of Regions

In the confusion matrix by region, we have the following data:

a. is the number of correct negative class predictions (actual negatives)
b. is the number of incorrect predictions of class positive (false positives)
c. is the number of incorrect negative class predictions (false negatives)
d. is the number of correct predictions of positive class (actual positives)
e. is the number of correct negative class predictions (actual negatives)
f. is the number of correct predictions of positive class (actual positives)
g. is the number of correct negative class predictions (actual negatives)

Construction of the Decision Tree

Decision trees are a classification model used in artificial intelligence, the main characteristic of which is its visual contribution to decision making.

A tree is graphically represented by a set of nodes, leaves, and branches. The central node or root is the attribute from which the classification process begins; the internal nodes correspond to each of the questions about the particular attribute of the problem. A child node represents each possible answer to the questions. The branches coming from each of these nodes are labeled with the possible values of the attribute. The end nodes or leaf nodes correspond to a decision, which coincides with one of the class variables of the problem to be solved (Witten & Frank, 2005).

The following figure shows the generation of the decision tree according to the data classification performed by the ID3 algorithm. (Figure 3)

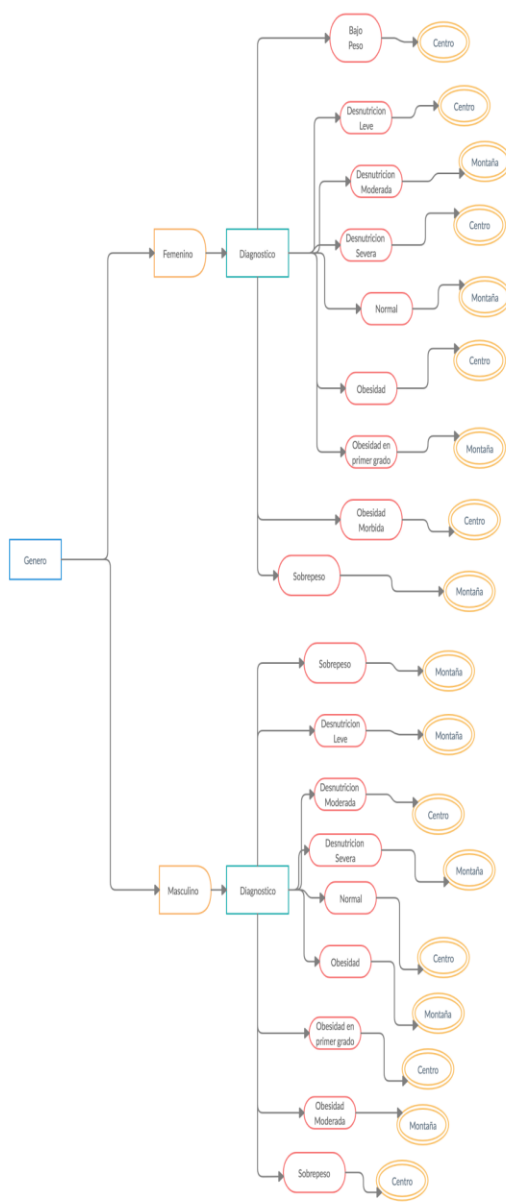

Figure 3 Decision tree generation (ID3 algorithm)
Source: Prepared by the authors

Results and Discussion

After collecting the information and implementing the ID3 algorithm for the construction of a decision tree, the graph shows two child nodes based on gender. In the case of Male, the following results are shown; the diagnoses Underweight (LW), Mild Malnutrition (MM), Obesity (OBD), and Morbid Obesity (OM), are concentrated with higher prevalence in men who live in the central zone of the state of Guerrero.

On the other hand, the preponderant diagnoses in men who live in the Mountain region are; Moderate Malnutrition (MD), Normal Nutrition (NR), Obesity (OB), and Overweight (SP). In the case of women, the data for underweight (LB), mild malnutrition (LD), severe malnutrition, obesity (OBD), and morbid obesity (OM) were found to be preponderant in the mountain region of the state of Guerrero.

The central and mountain regions have all the possible diagnoses defined by the WHO. From this information, we conclude that the mountain region is where the highest number of malnutrition is concentrated on women. However, in contrast to this data, the majority of men in this region suffer from grade one obesity. (See Figure 4).

Figure 4 Results by Region (State of Guerrero)
Source: Prepared by the authors

On the other hand, the possibility of increasing the data sample, as well as using other algorithms, methods, and techniques of artificial intelligence applied to the establishment of various diagnoses, could generate some relationships or patterns that would enrich the information obtained in this study. In the area of health, the countries consider it very important to be able to identify the factors that determine the prevalence and intensity of the different conditions.

In particular, the state of Guerrero faces the challenge of identifying these factors, which are strictly related to the educational level of the population, the socioeconomic situation, access to food, the quality of services associated with their housing, the environment, and the culture.

Conclusions

The experience of working with real data and a representative sample allows us to explore techniques in information processing that can generate and provide meaningful knowledge from patterns in the data. In this way, we investigated and highlighted the patterns by region of the nutrition of the inhabitants of the state of Guerrero. Using the WEKA tool, we were able to use data mining algorithms to visualize relationships or patterns in order to observe variables and attributes in a disaggregated way, such as the nutritional status of people living in the 7 regions of the state of Guerrero. In this analysis, we found some relationships that confirm the statistics of the WHO regarding the calculation of the BMI.

It is essential to mention that there are ailments that we did not initially imagine could occur in the mountain region of the state of Guerrero, such as overweight and obesity, since it is one of the poorest municipalities in the world (UN, 2007).

The study confirms that severe malnutrition exists in the central zone of the state of Guerrero and the mountain region. We also observed that men tend to be obese in the central zone of the state. That female patients tend to be obese in the first degree in the mountain region, contrary to some hypotheses due to the difficulty of food access in this region.

We were able to identify that the central and mountain regions are those where there is a higher rate of overweight in both men and women. A more exhaustive analysis of the relations between the variables may generate greater knowledge of the phenomenon studied.
The research I presented considered the analysis from the variables of gender, height, medical diagnosis, age, of a total of 17 variables that the system stores (employment, socioeconomic situation, blood pressure, blood type, family history, among others).

Given the above, we consider it essential to analyze the relationships among the other variables to understand, among other aspects, the direct and indirect causes that promote these chronic degenerative conditions. The use of these techniques could help us to visualize the characteristics by geographical region of these ailments, among other possibilities. They could also provide clarity about causal relationships that may be overlooked by health experts, allowing for better design of public policies in the field of preventive health.

References

BARQUERA, Simón et al. (2013)Prevalencia de obesidad en adultos mexicanos, 2000-2012. *Salud pública Méx* [online]. 2013, vol.55, suppl.2, pp.S151-S160. ISSN 0036-3634. Recuperado de: http://saludpublica.mx/index.php/spm/article/view/5111/10117

CONAPO (2010) (Consejo Nacional de Población). “Proyecciones de la Población e Indicadores Demográficos Básicos de México a Nivel Nacional 2010-2050. México: CONAPO, 2013”. Recuperado de: http: //www.conapo.gob.mx/es /CONAPO/ Proyecciones_de_la_Poblacion _ 2010-2050

Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016. Recuperado de: https://www.cs.waikato.ac.nz/ml/weka/citing.html

ENSANUT (2016) Encuesta Nacional de Salud y Nutrición de Medio Camino. Instituto Nacional de Salud pública Recuperado de: http:// ensanut.insp.mx/informes/ Guerrero-OCT.pdf

INEGI (2009) (Instituto Nacional de Estadística y Geografía).“Estadísticas Históricas de México”. México: INEGI, 2009. Recuperado de: http: //www.inegi.org.mx/prod_serv/contenidos/espanol/bvinegi/productos/ integracion/pais/historicas10/EHM2009.pdf

Instituto Nacional de Salud Pública (2016) (ENSANUT MC 2016) (2013-2018). “Encuesta Nacional de Salud y Nutrición: Evidencia para la Política Pública en salud”. Obtenida el 20 de noviembre de 2017, de https://www.gob.mx/cms/uploads/attachment/file /209093/ENSANUT.pdf

Gallardo-Bernal, Iván (2015). “Ataques Informáticos Basados en la Integridad de la Información”, Revista Salud y Administración Volumen 2 Número 5 .Consultado en septiembre de 2017 http://www.unsis.edu.mx/revista /doc/vol2num5/ A4_Ataques_Info.pdf

Larose, Daniel T. (2005). “Discovering Knowledge in Data an Introduction to Data Mining”. Hoboken, New Jersey. Jhon Wiley & Sons, Inc Publication. 222p

OCDE (2017) Obesity Update. Recuperado de: http://oment.uanl.mx/descarga/obesity-update-2017_ocde.pdf

OMS (2016) Informe de la Comisión para acabar con la obesidad infantil. Recuperado de: https://apps.who.int/iris/bitstream/handle/10665 /206450/9789243510064_spa.pdf;jsessionid=B17AC9A4B6590CF5A00622E85A667EAD?sequence=1

OMS (008) “Métodos poblacionales e individuales para la prevención y el tratamiento de la diabetes y la obesidad”. http://www.paho.org/hq/index.php?option=com_ docman&task=doc_ view&gid=15558&Itemid

ONU (2007) ONU: Cochoapa El Grande es el municipio más pobre de AL. En La Jornada (20/03/2007) Recuperado de: https://www.jornada.com.mx/2007/03/20/index.php?section=estados&article=031n2est

PAHO (2016a). Core Health Indicators in the Americas. Washington DC.

PROSESA (2013-2018) Programa Sectorial de Salud. México. Gobierno Federal. Reuperadode: https://www.gob.mx/salud/acciones-y-programas/programa-sectorial-de-salud-21469

Quinlan JR. (1996) “Learning Decision Tree Classifiers”. ACM Computing Surveys 1996; 28(1): 71-72.

Sánchez, C.L. (2013). “La información es el activo más valioso para muchas empresas, Sin embargo, no la tienen asegurada.” Disponible en: http://bts.inese. es

Witten, I. H., & Frank, E. (2005). “Data Mining: Practical machine learning tools and techniques”. San Francisco: Morgan Kaufmann Publishers.