ISSN 1870-4069

An Analysis of Demographic and Dietary Data with an Oral Health Approach: A Preliminary Study using Genetic Algorithms

Laura A. Zanella-Calzada[1], Carlos E. Galván-Tejada[1], Nubia M. Chávez-Lamas[2], María Del Carmen Gracia-Cortez[2], Jorge I. Galván-Tejada[1]

[1] Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Zacatecas, Mexico
{lzanellac, ericgalvan, gatejo}@uaz.edu.mx
[2] Universidad Autónoma de Zacatecas, Unidad Académica de Odontología, Zacatecas, Mexico
{nubiachavez, gacc005340}@uaz.edu.mx

Abstract. Oral health is one of the main components in the quality of life of people, since it is a determining factor in general health that affects the risk of suffering from other conditions, such as chronic diseases. Dental caries is the condition that most affects oral health worldwide, and occurs in about 90 % of people. The high prevalence in dental caries is caused by the diverse elements that interact simultaneously in their favor, such as the nutritional and socioeconomic elements. Based on this problem, this study proposes the analysis of a series of dietetic and demographic features compiled by the National Health and Nutrition Examination Survey 2013 – 2014, in order to obtain a model that allows the automatic classification of subjects according to their oral health status, presence or absence / restorations of caries. The methodology was carried out in three steps, starting with a data preprocessing, followed by a feature selection using a genetic algorithm and a final validation through a statistical analysis. The developed model is made up of five features (of the initial 188), and reached an area under the curve of 0.748, which is a statistically significant value. According to the results obtained, it was possible to conclude that the proposed model presents a preliminary result for the development of a low-cost support tool that helps the diagnosis of dental caries and the reduction of this condition.

Keywords: oral health, dental caries, classification, feature selection, genetic algorithm, statistical analysis.

1 Introduction

Chronic diseases is one of the main problems in public health and, nowadays, the pattern of diseases has changed, turning oral diseases into an important public health problem throughout the world. Oral diseases have a high incidence and

prevalence in all regions of the world, especially in disadvantaged and socially marginalized populations. According to the World Health Organization (WHO), the treatment of various oral conditions is extremely expensive and is not feasible in most low and middle income countries, being the fourth most expensive cause to treat [1]. Oral health presents an essential component in the quality of life due to its great influence on general health, since this condition can increase the risk of chronic diseases, such as: cardiovascular and cerebrovascular, diabetes mellitus and respiratory [2].

The most common condition in oral health is dental caries, which according to the WHO, affects between 60% and 90 % of children between five and 17 years. There is a series of determinants that favor this condition, such as carbohydrate consumption, food characteristics, plaque removal, among others, that interact simultaneously with variables that correspond to different orders of biological processes, complex historical-cultural structures, social relations, socioeconomic level, educational level, among others [3, 4].

Due to the difficulty of controlling the incidence of caries, which is caused by the large number of factors that influence, recent studies have implemented algorithms and performed analyzes based on computer-aided diagnosis (CADx) to develop prediction and classification models for preventive diagnosis and reduction of dental caries prevalence, looking for the main factors that affect this condition [5]. This study presents an analysis of two types of determinants that affect the important condition in oral health, dental caries. These determinants are demographic and dietary features that have been collected by the National Health and Nutrition Examination Survey (NHANES), 2013 – 2014. With a feature selection through a genetic algorithm, a multivariate model is developed that shows significant results in the classification of subjects with presence of this condition of subjects against absence, validated under a statistical analysis.

This work is organized as follows; Section 2 presents the methodology followed in three main steps, data preprocessing, feature selection and validation. In Section 3, the results obtained are shown and in Section 4 these results are discussed. Finally, conclusions are briefly described in Section 5.

2 Methodology

The methodology of this work is given in Figure 1, which is a flowchart of the stages that were followed for the development of this study.

A data preprocessing step (A) was performed to avoid any problem related to missing data or outliers that could affect the later stages. Then, in (B) a feature selection step is presented, which was carried out using the genetic algorithm "Galgo" to find the features that show the most significant behavior for the correct and automated classification of the subjects. Finally, (C) presents a validation step, where the features that were selected were subjected to a logistic regression, obtaining a general model for the classification of subjects that allows knowing the true positive and true negative rates through the calculation of

the receiver operating characteristic (ROC) curve and the area under the curve (AUC).

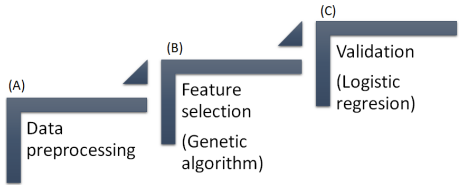

Fig. 1. Flowchart of the methodology followed.

All the methodology was performed using "R" (version 3.4.4) [6], which is a free software environment for statistical computing and graphics.

2.1 Data Description

NHANES is a program of studies designed to assess the health and nutritional status of children and adults in the United States of America (USA)and was founded by the Centers of Disease Control and Prevention (CDC) and the National Center for Health Statistics (NCHS). The surveys conducted by this program are unique, since they combine interviews and physical examinations [7]. NHANES collects information from different types of data and, in turn, this information is included in six main contexts; demographic, dietary, examination, laboratory, questionnaire and limited access. For this work, the demographic and dietary datasets were used;

- Demographic: it provides individual, family and household level information in different topics (income of households and families, size of households and families, pregnancy status, among others).
- Dietary: it provides detailed information on dietary intake, in order to estimate the types and amount of food and beverages consumed, in addition to estimating the intake of energy, nutrients, and other food components.

These datasets were contained by 188 features, and they were used as input features; while the condition of dental caries (absence or presence / restorations) was used as output feature. The subjects that were contained in those datasets belong to different counties in the USA and they were randomly selected with a computer algorithm by NHANES. The total number of subjects was 9812 (3690 controls / 6122 cases), 4982 females and 4830 males, and they were in a range age between zero and 89 years old.

2.2 Data Preprocessing

The two main purposes of the data preprocessing stage were to avoid problems of missing data and outliers.

Initially, all incomplete cases were manually removed, eliminating those subjects that presented $\geq$ 70 % of missing data. Of the rest of the subjects, all the missing values were imputed through the "nearestneighborimpute" function of the package "FRESA.CAD" (version 3.0.1) [8]. This function searches for any "NA" (not available) that is present in the dataset and looks for the row of complete observations that have the closest interquartile range normalized Manhattan distance to the rows that present missing values. In case that more than one row has similar minimum distances, the median value is used.

Then, the data was normalized using Z normalization, forcing the data to have a standard normal distribution. It was calculated with Equation 1, where x_i refers to the data, μ to the mean value and σ to the standard deviation:

$$Z_i = \frac{x_i - \mu}{\sigma}. \tag{1}$$

Finally, the singular values were removed, eliminating those columns that presented multiple values among themselves or the same value along the entire rows. This step was performed to avoid redundant or non-significant information.

2.3 Feature Selection

The feature selection was carried out through "Galgo" (version 1.1) [9], an R package for the selection of multivariate variable using genetic algorithms. In addition, Galgo presents a series of functions for the analysis of the population contained in the models and for the reconstruction and characterization of the models.

The Galgo procedure begins with a set of subsets of features or genes (named chromosomes) contained by data that was randomly selected. Then, each chromosome is evaluated based on its ability to predict a dependent variable, measuring its level of accuracy. The general idea is to replace the initial set of data with new data that conforms variants of chromosomes with a higher classification accuracy and repeat this process until achieving a desired level of accuracy. The progressive improvement of the chromosomes is carried out through a series of steps that resemble the process of natural selection, which consists of three steps; selection, mutation and crossover.

Looking for the increment of the solution space that is explored, independent chromosomes can evolve in isolated environments (named niches). Chromosomes can occasionally migrate between niches ensuring that that good solutions can recombine.

The classification methods that are included in Galgo are "k-nearest-neighbors", "nearest centroid", "support vector machines", "neural networks", "classification trees" and "discriminant functions" [10].

The main four steps that Galgo follows are briefly described.

1. **Setting-up the analysis.** This step is necessary to specify the input and the output features, the statistical model, the desired accuracy, the error estimation scheme and the parameters that decide the search environment of

the genetic algorithm. It is important to mention that the estimation of the error can be defined in two levels; a validation strategy for training / testing, and within the training process using k-fold cross-validation, random splits or re-substitution.

2. **Searching for relevant multivariate models.** Starting with a random set of chromosomes, the genetic algorithm procedure will look for a diverse collection of statistically significant local solutions. A sufficiently large number of chromosomes must be selected to have a good representation of the solution space.
3. **Refinement and analysis of the set of selected chromosomes.** The chromosomes that were selected have a length that was previously defined. Although the models have reached the desired classification accuracy, a refinement step is carried out, since there is a possibility that not all the model genes have a significant contribution to accuracy. Through a backward selection strategy, only those genes that effectively contribute to the classification task are selected.
4. **Development of a representative statistical model.** Finally, in this part of the analysis a single representative model of the selected chromosomes is obtained. For this step, a forward selection strategy was implemented based on the step-wise inclusion of the most frequent genes in the chromosome.

For this work, the settings used in this experimentation were the following: chromosomes composed of five genes, with 200 evolutionary processes in 200 big bangs. These settings were selected given the number of features and patients to avoid overfitting.

On the other hand, the classification model used was the nearest centroid, which is a supervised method that assigns to the observations the class label of the training samples whose mean or centroid is the closest to the observations. For each set of data points belonging to the same class, the centroid vectors are calculated. If there are k classes in the training set, there are k centroid vectors. The test samples are classified in the class with the nearest centroid.

For the training procedure, given the labeled training samples $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$, with class labels $y_i \in Y$, the class centroids $\boldsymbol{\mu}_k$ are calculated with Equation 2, where C_k is the set of indices of samples belonging to class $k \in Y$:

$$\boldsymbol{\mu_k} = \frac{1}{|C_k|} \sum_{l \in C_k \boldsymbol{x}_i} . \tag{2}$$

The prediction function $\hat{y}$, which is the class assigned to an observation $\boldsymbol{x}$ is calculated with Equation 3:

$$\hat{y} = min_{k \in Y} ||\boldsymbol{\mu}_k - \boldsymbol{x}||. \tag{3}$$

2.4 Validation

The validation stage was carried out in order to evaluate the multivariate model resulting from the feature selection. Initially, the model was subjected to a

Logistic Regression (LR) to obtain a general model for the classification of subjects using the set of selected features. LR is an analysis that consists of a statistical technique to model the relationship between the features. This method belongs to the statistical methods where the contribution of different factors in the occurrence of a simple event is measured. The main objective of LR is to model the influence of the probability of an event. The simplest representation of a model obtained by this method is presented in Equation 4, where y from $logit(y = 1)$ is the dependent variable or the outcome feature that is necessary to be subjected to a logarithmic transformation ($logit$) because the initial equation of the model is of exponential type and by this transformation it is possible to use it as a lineal function, w is an offset term that can be included, β_1 is the slope and x is the independent variable or the analyzed feature. Models can be composed by the number of independent variables needed [12]:

$$logit(y = 1) = w + \beta_1 x. \tag{4}$$

On the other hand, the AUC value is a standard method used for the evaluation of the accuracy obtained by the model, and is calculated with the relationship between the specificity and the sensitivity [13].

The sensitivity parameter is referred to the proportion of data that belongs to a condition and it is classified as positive. This value is obtained with Equation 5, where TP represents the quantity of true positives and FP represents the quantity of false positives:

$$PPV = \frac{TP}{TP + FP}. \tag{5}$$

The specificity parameter is referred to the proportion of subjects without a condition that are classified as negative. This value is obtained with Equation 6, where TN represents the quantity of true negatives and FN represents the quantity of false negatives:

$$NPV = \frac{TN}{TN + FN}. \tag{6}$$

3 Results

The results obtained from the methodology performed are presented in this section.

From the preprocessing step, a series of features were eliminated according to the missing data and singular values that were present in the dataset, maintaining a total of 136 dietary and demographic features. Then, those features were submitted to the Galgo algorithm for a feature selection.

Figure 2 presents a graph where all the features are ordered according to their degree of appearance in the chromosomes throughout the iterations. The features shown in black have the highest frequency, while the features shown in gray have the lowest.

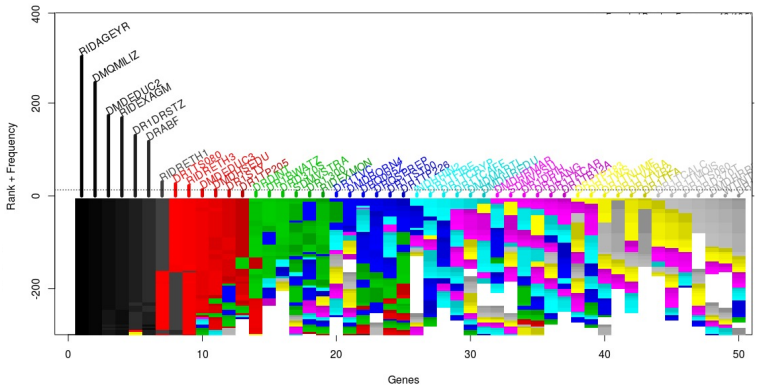

Fig. 2. Graph of the frequency with which the features appear on chromosomes, ordered descending. Vertical axis presents the frequency number and horizontal axis presents the chromosomes.

Figure 3 presents the performance graph of forward selection, showing the classification accuracy achieved for each feature included in the model. The features are arranged according to their frequency of appearance, from the highest to the lowest.

After the forward selection step, the backward elimination process was performed, obtaining a final model that presented the best accuracy without any redundant information. This model was contained by the five features that are described in Table 1.

Table 1. Features obtained from the backward elimination step.

Feature	Description
RIDAGEYR	Age in years, at the time of the screening interview, is reported for survey participants between the ages of 1 and 79 years of age.
RIDEXAGM	Age in months of the participant at the time of examination.
DMQMILIZ	Have you ever served on active duty in the U.S. Armed Forces, Military reserves, or National Guard?
DMDEDUC2	What is the highest grade or level of school you have completed or the highest degree you have received?
DR1DRSTZ	Dietary recall status.

Finally, after the feature selection, the validation step was carried out, where the ROC curve presented in Figure 4 was obtained, having an AUC of 0.748.

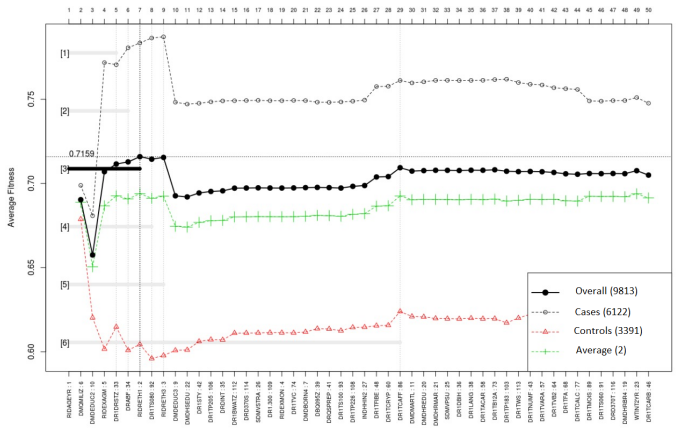

Fig. 3. Graph of the forward selection step. Vertical axis presents the accuracy value and horizontal axis presents the chromosomes ordered according with their frequency appearance.

4 Discussion

The results obtained show a multivariate model, contained by demographic and dietary features, which was obtained through the Galgo genetic algorithm, which presents a statistically significant performance in the classification of subjects who show absence of dental caries from those who have presence or restorations of them.

Figure 2 presents the graph of the frequency that each feature obtained according to the number of times they appeared in the different chromosomes that developed throughout the Galgo process, where it is possible to observe that from the total features, there where seven with a significant number of frequency, which are those presented in black; RIDAGEYR, DMQMILIZ, DMDEDUC2, RIDEXAGM, DR1DRSTZ, DRABF and RIDRETH1. Then, in Figure 3 the graph of the forward selection step is observed, where the accuracy obtained for each feature that is included in the model is presented; when the model is contained by the seven most frequent features, the best accuracy is achieved.

In addition, it is possible to observe that the classification of the cases (-○-) is much more accurate than the classification of the controls (-△-), making that the overall (-●-) classification reduces its accuracy. This may occur due to the low significant information that the features may be presenting for the control subjects or to the number of controls presented in the data set, which can be resolved by increasing the number of subjects and / or features.

Then, Table 1 includes the five features that were selected through the backward elimination step, eliminating two of the features that were selected in the previous step; DRABF and RIDRETH1. These features were removed because they didn't present any significant information for the classification of subjects or their contribution was redundant.

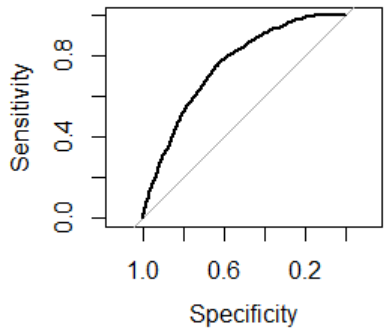

Fig. 4. ROC curve obtained from the classification of subjects.

The information presented in the multivariate model is related to the age of the subjects, the educational level, the state of the diet and the duty of activity in USA services. As mentioned above, one of the main factors that affect the state of oral health is age, with young people being the most affected by dental caries. On the other hand, the state of dietary recall is used to know the individual foods and the total intake of nutrients, which indicates the quality and integrity of the subjects, being one of the determinants of dental caries that was also mentioned above.

The educational level and the participation in the services of the USA are features related to the socioeconomic level. The educational level is significant because the low income regions have less educational opportunities; while participation in the services of the USA is an important feature taking into account that the main objective population of NHANES is the non-institutionalized civilian resident population of the USA, which may be people who decided to emigrate in search of better economic opportunities and one of the requirements for the Armed Forces of the USA, the Military Reserves and the National Guard is being a citizen or a permanent resident. Therefore, it is possible to justify the selection of those features for the developed model.

Finally, Figure 4 presents the ROC curve obtained according to the true positives and true negatives calculated with the developed model; its AUC value was 0.748, which is statistically significant since it means that from the total subjects, 74.8 % were correctly classified.

5 Conclusions

This paper presents the analysis of a series of demographic and dietary features, in order to develop a model that can present a tool for specialists in the preventive diagnosis of dental caries and the reduction of the incidence of this condition. This analysis was performed in three main stages, data preprocessing, feature

selection and validation. In the stage of feature selection, a multivariate model was developed that contained the features that provided the most significant information in the classification of control and case subjects, while the validation stage allowed the evaluation of this multivariate model, obtaining statistically significant results.

Therefore, the model proposed in this work may represent a low-cost preliminary tool that helps in the diagnosis of dental caries and in the possible prediction of them, both for high and low income regions, to reduce their high incidence and avoid the high cost treatments that this condition represents.

References

1. World Health Organization (NCHS): Oral Health, http://www.who.int/oral-health/disease-burden/global/en/. Last accessed 05 June 2018
2. Ridao Marín, D.: Desarrollo de un Sistema de Ayuda a la Decisión para Tratamientos Odontológicos con Imágenes Digitales. Universidad de Málaga: Málaga, Spain, 10–12 (2017)
3. Espinoza Solano, M.; León-Manco, R.A.: Prevalencia y experiencia de caries dental en estudiantes según facultades de una universidad particular peruana. Rev. Estomatol. Hered. 25(3), 187–193 (2015)
4. Acuña Aguilar, L.D., Porras Cerón, D., Ríos Rueda, L.D.: Prevalencia de Lesiones Cariosas y Factores Asociados Presentes en Pacientes con SíNdrome de Down en las Fundaciones Fundown y san Luis Guanella de Bucaramanga. Universidad Santo Tomás: Bucaramanga, Colombia, pp. 12–16 (2017)
5. Gispert Abreu, E.D.L.Á., Castell-Florit Serrate, P., Herrera Nordet, M.: Salud bucal poblacional y su producción intersectorial. Rev. Cubana Estomatol.52, 62–67 (2015)
6. R: A Language and Environment for Statistical Computing, https://www.R-project.org/. Last accessed 17 June 2018
7. National Health and Nutrition Examination Survey Data. 2013–2014, http://www.cdc. gov/nchs/nhanes.htm. Last accessed 17 June 2018
8. FRESA.CAD: Feature Selection Algorithms for Computer Aided Diagnosis, https://CRAN.R-project.org/package=FRESA.CAD. Last accessed 17 June 2018
9. GALGO: an R package for multivariate variable selection using genetic algorithms, http://bioinformatica.mty.itesm.mx/galgo2. Last accessed 17 June 2018
10. Trevino, V., Falciani, F.: GALGO: an R package for multivariate variable selection using genetic algorithms. Bioinformatics 22(9), 1154–1156 (2006)
11. Das, T.: Machine Learning algorithms for Image Classification of hand digits and face recognition dataset. Machine Learning 4(12), 640–649 (2017)
12. Montgomery, D. C., Peck, E., Vining, G.: Introduction to linear regression analysis. John Wiley & Sons, Location (2015)
13. Berlanga Silvente, V., Vilà Baños, A.: AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography 17(2), 145–151 (2008)