A Preliminary Approach for Finding Linked Determinants between Diabetes and Caries using Genetic Algorithms

Laura A. Zanella-Calzada[1], Carlos E. Galván-Tejada[1], Nubia Chavez-Lamas[2], Miguel A. Cid-Baez[1], M. del Carmen Gracia-Cortés[2]

[1] Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Zacatecas, Zac, Mexico
{lzanellac, ericgalvan, drcidbaez}@uaz.edu.mx

[2] Universidad Autónoma de Zacatecas, Unidad Académica de Odontología, Zacatecas, Zac, Mexico
{nubiachavez, gacc005340}@uaz.edu.mx

Abstract. Two of the public health conditions that continue having an increase in their incidence and prevalence worldwide are dental caries and diabetes mellitus. Dental caries is a multifactorial disease that affects the hard tissues of the tooth, being the most widespread disease in man; while diabetes mellitus is a multifactorial chronic disease that involves a group of disorders that alter the production and use of insulin, being of the main causes of disease in high, medium and low income countries. In this work is proposed the use of genetic algorithms in order to select the most significant features from a database of demographic data obtained from the NHANES $2013 - 2014$, that best describe the difference between cases (diabetes and caries) and controls (healthy), according to their health status. After the feature selection, it is performed a validation stage, evaluating the multivariate model contained by these features. Within the results achieved shows a fitness > 0.980 and an $AUC = 0.975$, validating that the demographic features selected in this paper present a highly significant capacity to classify subjects that present caries and diabetes simultaneously from healthy subjects.

Keywords: dental caries, diabetes mellitus, health status, genetic algorithm, feature selection, multivariate model.

1 Introduction

There are currently 425 million adults with Diabetes mellitus (DM) in the world, that is, one in 11 adults, 46.5% of whom do not know they have it, and it is estimated that by 2040, this figure will increase to 642 million, with 5.1 million people dying as a direct result of Diabetes [5]. Mexico with 12.03 million people with diabetes ranks 6th worldwide [6] and the 1st. place in obesity worldwide [3]. Most blood glucose control is poor [9] coupled with low adherence to treatment and risk co-morbidities (such as cholesterol and high blood pressure) [1]. The

dramatic increase in the global prevalence of diabetes has led to an inevitable increase in biopsychosocial complications [25]. Mexico is a middle-income country, where a medium-low socioeconomic and educational status predominates (low human development index), it conditions an unfavorable health status (2 to 4 times higher, against a high status) [26]. This has a negative impact on the incidence of co-morbidities of oral pathologies, which are considered public health care, such as caries. [15].

Caries is defined as a preventable disease of chronic multifactorial origin that affects the hard tissues of the tooth, with a prevalence of 94% of individuals worldwide [27]. People with diabetes have an increased risk of gum inflammation (periodontitis) or gingival hyperplasia if their blood glucose is not properly managed. Periodontitis is a major cause of tooth loss and is associated with an increased risk of cardiovascular disease[18]. Other diabetes-related mouth conditions include dental caries 91.4%, gingivitis 97.1%, periodontitis 85.7%, oral candidiasis 17.1%, sensorineural disorders (burning mouth), salivary dysfunction and xerostomia 62.9%, and taste alteration 40%, periodontal abscesses 40% [12]. People with diabetes have a higher risk of developing periodontal disease than people without diabetes [10]. The prevalence of periodontal disease is more common among people with diabetes (92.6%) than among people without diabetes. This complication may be caused by chronic immunosuppression, delayed healing, or salivary hypofunction. The use of a toothbrush and a strict dental maintenance program is important in long-term oral health and for the prevention of complications. Regular oral check-ups provide a guarantee for early diagnosis, especially among previously undiagnosed diabetics and rapid management of any complications among patients with diabetes. On the other hand, oral health is considered to be an integral and essential part of overall health, and this can compromise quality of life. The phenomena of the binomial oral health and diabetes are highly complex, since they interact simultaneously and with variables that correspond to different orders: from biological processes, to complex historical-cultural structures and social relations.

Due to the difficulty that represents the control in the incidence of these conditions, caries and diabetes, which are caused by the large number of factors that influence, recent studies have implemented algorithms and performed analysis based on computer-aided diagnosis (CADx) in order to develop prediction and classification models for the preventive diagnosis and prevalence reduction of them looking for the main factors that affect this condition [21].

This paper is organized as follows, in section 2 is presented a description of the data used for this research and the methods that were carried out. Results and discussion obtained from the genetic algorithm process and the validation step are shown in section 3. Finally conclusions are described in section 4.

2 Materials and Methods

In this work is presented the development of a multivariate model contained by demographic features in order to classify subjects that present two different

conditions, the presence of caries and diabetes or the absence of caries and diabetes.

In this section is described the data from the National Health and Nutrition Examination Survey (NHANES) 2013 – 2014 that were used for this work, as well as the data preprocessing, feature selection and validation of the results.

Fig. 1. Flowchart of the methodology followed. A) Dataset used from the NHANES 2013 – 2014, B) Data preprocessing, C) Feature selection using a genetic algorithm, D) Validation through statistical analysis.

In Figure 1 is presented the methodology followed for the development of this research. Firstly, in (A) is shown the database used, which specifically corresponds to demographic data; then, in (B) is presented a preprocessing step where it was necessary to prepare the data in order to obtain a correct analysis of the features. In (C), the data were subjected to a feature selection through a genetic algorithm, presenting the features that shown better overall performance and individually for the classification of subjects in the two conditions mentioned before. Finally, in (D) is carried out a validation step, trough a statistical analysis, obtaining the Receiver operating characteristic (ROC) curve and the Area under the curve (AUC), with the purpose of evaluating the results obtained from the feature selection.

2.1 Data Description

The NHANES is a national program that design studies to perform a survey to assess the health and nutritional status of adults and children in the United States, including all ethnic groups. The survey combines interviews and physical examinations, allowing to develop studies using clinical, para-clinical and demographic characteristics (features) of individuals. NHANES is a program founded by the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the United States.

Content Description NHANES survey include several types of interviews, to cover a wide range of features, including demographic, socioeconomic, dietary, and health-related questions, described in detail in 1. One examination component that is critical to this study is dental care examination, as well as the

health-related question about the condition of diabetes, which are carried on by trained medical personnel.

In Table 1 is presented the description of the main datasets used in this work, which correspond to Demographic, Examination and Questionnaire.

Table 1. NHANES Data Description

Questionnaire Type	Description
Demographic	The demographics file provides individual, family, and household level information.
Examination	Public health significance in areas of surveillance, prevention, treatment, dental care utilization, health policy, evaluation of Federal health programs.
Questionnaire	information on: Acculturation, Alcohol Use, Diabetes Status, Health Insurance, Income.

Meta Data Health and Nutrition Examination Surveys are comprised by interviews applied to 27,631 subjects. These subjects were submitted to a series of different questionnaires, related to the different features. These subjects were randomly selected with a computer algorithm by NHANES. The algorithm consists of a complex multistage probability design to choose the participants, with a series of stages.

The subjects selected were women and men that belong to different counties in the USA. These counties were divided into 15 groups according to their characteristics, and from each group one county was randomly selected, obtaining the 15 counties for the surveys. Then, a sampling of segments was performed to each county, selecting between 20 and 24 segments, where from all the houses that correspond to the segments, a sample of about 30 households were selected within each each segment. Finally, the sampling of persons was carried out where the NHANES interviewers went to the selected households and asked for the information of the surveys, randomly selecting the members of the households that were part of the final database.

The main target population for NHANES is the non-institutionalized civilian resident population of the USA. The design of the population selection is based on the sampling of a larger number of specific subgroups that present particular public health characteristics, in order to increase the reliability and precision of health status indicators. NHANES started these design changes in 2011, including in its population the oversampled subgroups survey cycle:

– Hispanic persons;
– Non-Hispanic black persons;
– Non-Hispanic Asian persons;
– Non-Hispanic white and other* persons at or below 130% of the poverty level; and
– Non-Hispanic white and other* persons aged 80 years and older.

The subjects contained in the Demographic database, which was the used for this work, were 9,801 (male = 4,826, female = 4,976), and the total of features were 39. Subjects were located in an age range from 0 to 80 years old.

2.2 Data Preprocessing

The data preprocessing consisted on prepare the data for their correct analysis. Initially, it was necessary to unify the database, keeping those subjects that presented the information of the features that made reference to the dental caries and diabetes status.Those features were mixed in one, which was the feature used as output, where subjects that presented a positive condition in both were classified as '1' and subjects that presented a negative condition in both were classified as '0'.

Then, there were removed those features that presented $\geq 70\ \%$ of missing data. The remaining data were imputed, normalized and finally, those features that presented singular values after the normalization were removed.

The imputation of the data was performed through the 'rfimpute' function, from the 'randomForest' package (version 4.6-12, 2011-10-18) [14], for R. This function imputes the data replacing all the missing data with the median value of the column, then, the 'randomForest' function is called with the complete data to update the imputation of the missing values.

On the other hand, for the normalization step was used Equation 1, where x represents the value that needs to be normalized, μ represents the mean of the column and σ represents the standard deviation of the column:

$$Z = \frac{x - \mu}{\sigma}. \tag{1}$$

Features that presented singular values were those that had the same value in the whole column.

2.3 Feature Selection

For the feature selection, the data were subjected to an implementation of a genetic algorithm called 'Galgo' [24], which is a package for R.

Galgo performs a procedure for selecting models with a high fitness value and implements a series of functions for the analysis of the population of the selected models, besides the reconstruction and characterization of representative summary models [13].

To begin, from a random population of feature subsets of a given size, which are defined as chromosomes, an assessment for their ability to predict a dependent feature is performed to each one, obtaining a certain level of accuracy. The general principle is to replace the initial population with a new population that includes some variants of chromosomes with higher classification accuracy and to repeat the process enough times to achieve a desired level of accuracy. The continuous improvement of the chromosome population is driven under the principle of the natural selection, through three main steps, selection, mutation and crossover. Besides, in order to increase the proportion of the solution space, independent populations can be evolved in isolated environments, defined as niches, and chromosomes can migrate from one niche to another ensuring that particularly good solutions can recombine.

In summary, Galgo presents features as Gene objects, which are part of a Chromosome object or a set of n features that are included in a multivariate model that is evaluated through a fitness function, using a classification method as cost function. Then, a Niche object organizes chromosomes in populations and a World object is content of several niches. These objects are later ranked in a Galgo object, implementing the general genetic algorithm evolutionary process, saving the best chromosome. Finally, a BigBang object is developed, collecting the result of several iterations or searches for the further analysis. In this process are included most common genetic algorithms methods as reproduction, mutation, crossover, migration and elitism [24].

All the objects in Galgo have properties that have the option to users to control the process, besides they can add custom defined properties in order to add new functionality.

The parameters selected for this process were a choromosome size of five features, a total number of solutions of 300 and a goal fitness of 0.75.

The classification methods that are included in Galgo are "K-nearest-neighbor", "nearest centroid", "support vector machines", "neural networks", "classification trees" and "discriminant functions".

For this work the classification methods used was K-nearest-neighbor (K-NN), which is a nonparametric machine learning algorithm that has been widely used in pattern recognition and data mining. Its main purpose is to classify an unknown sample based on the known classification of its neighbors. This method works being fed with a training set that is contained with labeled samples that is used to classify objects. These objects are classified based on the K parameter, which means that they are assigned to a class that results to be the most similar among its closest K neighbors. In K-NN, it's firstly necessary to find the distance between the training and test data, one of the most used distance measure is the Euclidean distance, which can be calculated with Equation 2, where X is the training data with specific parameters (x_1 to x_n), and Y is the test data with specific parameters (y_1 to y_n) [8]:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

(2)

Then, the distances are sorted for all the objects in the training set and the nearest neighbor is found according to the minimum distance. The K parameter is the number of the closest neighbors in the place of interest and it can be obtained by the n-fold cross-validation method.

2.4   Validation

The validation consisted on submitting the multivariate model obtained from the feature selection into a statistical analysis, in order to evaluate its performance in the classification of subjects using the set of features selected.

Initially, a logistic regression (LR) is applied to the multivariate model, obtaining the general model that best classifies the subjects. LR is a statistical

analysis technique that consists in the modeling of the relationship between a dependent variable and an independent variable or a set of them. This technique is part to the statistical methods where is measured the contribution of different factors in the occurrence of a simple event. Thus, the main objective of LR is to model the influence of the probability of an event. In Equation 3 is presented the simplest representation of a model, being y from $logit(y = 1)$ the dependent variable, which represents the outcome feature, that is subjected to a logarithmic transformation ($logit$). This transformation is performed because the initial equation of the model is exponential and through this step the model becomes a lineal function. w is an offset parameter that can be included or not, β_1 is the slope and x is the independent variable or the input feature. Models can be content by n number of independent variables [16]:

$$logit(y = 1) = w + \beta_1 x. \tag{3}$$

Then, the general model that was obtained through the LR technique was evaluated measuring its true positives - true negatives proportion, obtaining its ROC curve and the AUC value. For these measurements, the $pROC$ (version 1.12.1, 2018-05-06) [20] package for R was used.

The AUC value is a standard method used for evaluating the accuracy of the model, and is calculated measuring the relationship between the specificity and the sensitivity [4].

Sensitivity is the proportion of data that belongs to a condition and it's classified as positive. This parameter is calculated with Equation 4, where TP is the quantity of true positives and FP is the quantity of false positives:

$$PPV = \frac{TP}{TP + FP}. \tag{4}$$

Specificity is the proportion of data that doesn't belong to a condition and it's classified as negative. This parameter is calculated with Equation 5, where TN is the quantity of true negatives and FN is the quantity of false negatives:

$$NPV = \frac{TN}{TN + FN}. \tag{5}$$

Finally, a comparison of the ROC curve and AUC value obtained between a model using the 39 initial features and the model of the five features selected was performed in order to do a cost benefit analysis of the results achieved.

All the data analysis were realized with the free software R (version 3.3.1) [19].

3 Results and Discussion

In this section the results obtained for each stage of the methodology are exposed, as well as the discussion.

Data preprocessing In the data preprocessing step, the 9,801 subjects were reduced to 3,552 (male = 1,812 , female = 1,740 / controls = 3041, cases = 511)

and the 39 demographic features were reduced to 32. The great reduction of the subjects is because most of them didn't have their complete information of their dental caries or diabetes status, besides, only the subjects that presented a positive result for both conditions were preserved, as well as the subjects that presented a negative results for both conditions. On the other hand, the reduction of the features is related to those features that presented singular values or a high number of missing data. The relationship between controls and cases may look disproportionate, however, the quantity of data for both conditions was enough to reach a significant performance in the classification of subjects.

Feature selection From the feature selection step, Galgo developed a graph of the gene rank and frequency, shown in Figure 2, with which the features appeared in the different chromosomes that were constructed by the iterative process of the genetic algorithm. In this graph, those features that were present in more chromosomes are in black while the features that were present in less are in gray, being possible to observe that the most frequent features were 'RIDEXAGM', 'RIDAGEYR', 'DMDEDUC3', 'DMQMILIZ'.

In Figure 3, it is presented a graph of the performance of fitness throughout the iterations or generations of the Galgo process, where is possible to observe that the fitness reached a significant value of around 0.99, which means that the classification of subjects using the multivariate model created is highly accurate, being possible to classify subjects in the two conditions using a reduced number of demographic features.

Finally, when Galgo finished its iterations, a forward selection and backward elimination were performed, obtaining a model contained by five features, shown in the heat map of Figure 4. This model was the result of the genes combination between the different chromosomes and the reduction of redundant or not significant information, where is shown that the strongest feature is 'RIDEXAGM', followed by 'DMDHSEDU', 'DMDEDUC3', 'DMQMILIZ' and 'RIDAGEYR'. From the final model, four of its features presented the highest frequency according to the graph of Figure 2, and even when the feature 'DMDHSEDU' didn't present a significant frequency of appearance, which means that is not presenting an important contribution for the chromosomes, it may be presenting good behavior together with the rest of the features, being the second strongest feature.

The features of the best chromosome are described in Table 2, exposing that two are referred to the age of the subjects, two are referred to the educational level and one is referred to the service of the country. This means that the age of the subjects represents important information in the presence or the absence of dental caries and diabetes simultaneously, since the age is identified as one risk factor in dental caries as well as in diabetes [11, 22]. On the other hand, the educational level and the participation in the U. S. services are features related with the socioeconomic level, which is one the determinants of both conditions too. Educational level is significant because low-income regions have fewer educational opportunities; while the participation in the U.S. services is a significant feature taking into account that the main target population

for NHANES is the non-institutionalized civilian resident population of the U.
S., which may be people that decided to migrate looking for better economic
opportunities and one of the requirements for the U.S. Armed Forces, Military
reserves and National guard is being a citizen or a permanent resident. Taking
into account that people that don't have access to a good educational level are
not enough informed about the care to avoid developing these conditions, like
learning to have a good oral hygiene, the knowledge that health programs can
offer, the eating habits, the diet quality, among others, besides possibly not
having access to a health system, it becomes difficult to be under a treatment
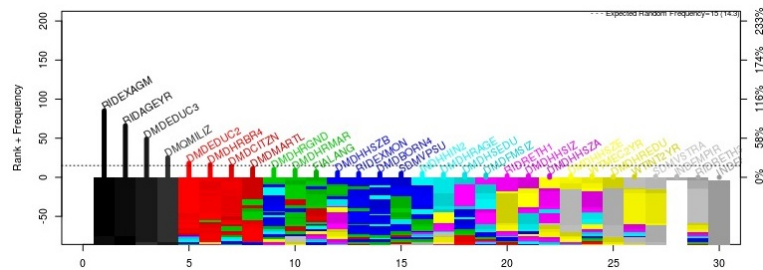that allows to combat these conditions [17, 2, 7, 23].

Fig. 2. Graph of the gene rank and frequency of appearance of the features in the
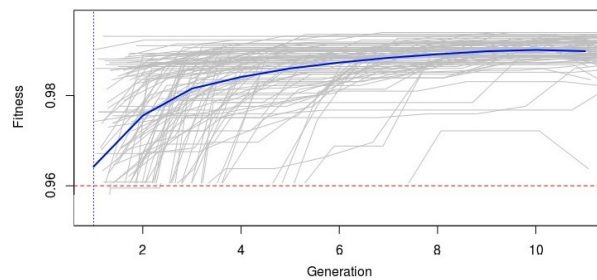developed chromosomes.

Fig. 3. Graph of the performance of fitness throughout the iterations of Galgo.

Validation Finally, in Figure 5 are shown the ROC curves that were obtained
with the general model using the 39 initial features (A) and with the general
model using the features selected (B), obtaining an AUC value of 0.979 and
0.975, respectively, which are statistically significant. As is possible to observe

both curves present a very similar performance, as well as the AUC value, which means that the features the are contained in the model are those that present the most significant contribution in the classification of subjects.
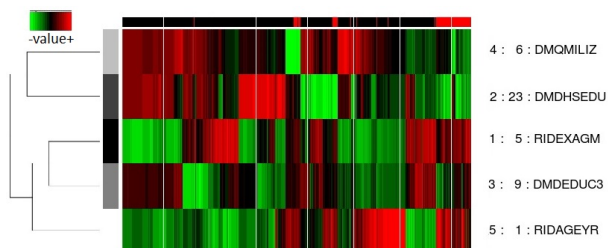

Fig. 4. Heat map of the best chromosome reached by the Galgo process.

Table 2. Best chromosome obtained from the feature selection.

| Questionnaire Type	Description
RIDEXAGM	Age in months of the participant at the time of examination.
DMDHSEDU	HH reference person's spouse's education level.
DMDEDUC3	Highest grade or level of school completed or the highest degree received.
DMQMILIZ	As SP ever served on active duty in the U.S. Armed Forces, military Reserves, or National Guard?
RIDAGEYR	Age in years of the participant at the time of screening.
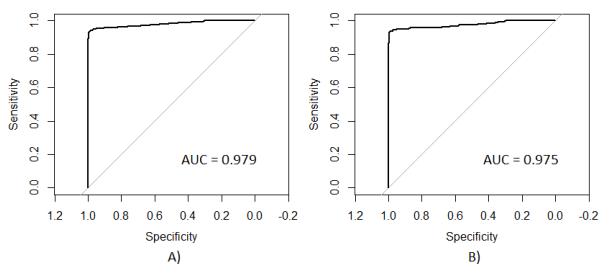
Fig. 5. ROC curves obtained with the modeling of (A) the initial 39 features and (B) the features selected.

4   Conclusion

This paper presents the analysis of a demographic features set in relationship with dental caries and diabetes, in order to develop a model that can present a tool for specialists in the preventive diagnosis of these conditions and the reduction of the incidence of them. This analysis was performed in three main stages, data preprocessing, feature selection and validation. In the feature selection stage was developed a multivariate model contained by features that provided the most significant information in the classification of control and case subjects, since they are important risk factors for these conditions according to literature; while the validation stage allowed to evaluate this multivariate model, obtaining statistically significant results.

Therefore, the model proposed in this work can represent a preliminary tool of low cost that helps in the diagnosis of dental caries and diabetes together, besides make a possible prediction of them, having as main objective the regions of low socio-economic and educational level, in order to reduce their high incidence and avoid the high cost treatments that represents these conditions.

References

1. Alegre-Díaz, J., Herrington, W., López-Cervantes, M., Gnatiuc, L., Ramirez, R., Hill, M., Baigent, C., McCarthy, M.I., Lewington, S., Collins, R., et al.: Diabetes and cause-specific mortality in mexico city. New England Journal of Medicine 375(20), 1961–1971 (2016)
2. Araujo, D., Marquezin, M., Barbosa, T., Fonseca, F., Fegadolli, C., Castelo, P.: Assessment of quality of life, anxiety, socio-economic factors and caries experience in brazilian children with overweight and obesity. International journal of dental hygiene 15(4), e156–e162 (2017)
3. Barriguete-Meléndez, J., Hercberg, S., Galán, P., Parodi, A., Baulieux, J.: Nutrinet-salud méxico. prospective study online: 2018-2028. relationship between nutrition and health and determinants of dietary habits and nutritional status. Cirugia y cirujanos 86(1), 8–19 (2018)
4. Bourne, D.A.: Mathematical modeling of pharmacokinetic data. Routledge (2018)
5. Cho, N., Shaw, J., Karuranga, S., Huang, Y., da Rocha Fernandes, J., Ohlrogge, A., Malanda, B.: Idf diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes research and clinical practice 138, 271–281 (2018)
6. Dávila-Torres, J., González-Izquierdo, J.d.J., Barrera-Cruz, A.: Obesity in mexico. Revista Médica del Instituto Mexicano del Seguro Social 53(2), 240–249 (2015)
7. Elamin, A., Garemo, M., Gardner, A.: Dental caries and their association with socioeconomic characteristics, oral hygiene practices and eating habits among preschool children in abu dhabi, united arab emirates—the noplas project. BMC oral health 18(1), 104 (2018)

8. Fadaei-Kermani, E., Barani, G., Ghaeini-Hessaroeyeh, M.: Drought monitoring and prediction using k-nearest neighbor algorithm. Journal of AI and data mining 5(2), 319–325 (2017)

9. Flores-Hernández, S., Saturno-Hernández, P.J., Reyes-Morales, H., Barrientos-Gutiérrez, T., Villalpando, S., Hernández-Ávila, M.: Quality of diabetes care: the challenges of an increasing epidemic in mexico. results from two national health surveys (2006 and 2012). PloS one 10(7), e0133958 (2015)

10. Hugoson, A., Thorstensson, H., Faltt, H., Kuylenstierna, J.: Periodontal conditions in insulin-dependent diabetics. Journal of clinical periodontology 16(4), 215–223 (1989)

11. Hunter, P.B.: Risk factors in dental caries. International dental journal 38(4), 211–217 (1988)

12. Lalla, E., Cheng, B., Lal, S., Kaplan, S., Softness, B., Greenberg, E., Goland, R.S., Lamster, I.B.: Diabetes mellitus promotes periodontal destruction in children. Journal of clinical periodontology 34(4), 294–298 (2007)

13. Li, L., Weinberg, C.R., Darden, T.A., Pedersen, L.G.: Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method. Bioinformatics 17(12), 1131–1142 (2001)

14. Liaw, A., Wiener, M.: The randomforest package. R News 2(3), 18–22 (2002)

15. Mendoza Romo, M.Á., Padrón Salas, A., Cossío Torres, P.E., Orozco, M.S.: Prevalencia mundial de la diabetes mellitus tipo 2 y su relación con el índice de desarrollo humano. Revista Panamericana de Salud Pública 41, e103 (2018)

16. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to linear regression analysis. John Wiley & Sons (2015)

17. Oyedele, T., Fadeju, A., Adeyemo, Y., Nzomiwu, C., Ladeji, A.: Impact of oral hygiene and socio-demographic factors on dental caries in a suburban population in nigeria. European Archives of Paediatric Dentistry 19(3), 155–161 (2018)

18. Papapanou, P.N.: Periodontal diseases: epidemiology. Annals of periodontology 1(1), 1–36 (1996)

19. Ripley, B.D.: The r project in statistical computing. MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network 1(1), 23–25 (2001)

20. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Müller, M.: proc: an open-source package for r and s+ to analyze and compare roc curves. BMC Bioinformatics 12, 77 (2011)

21. Singh, P., Pal, A.B., Anburajan, M., Kumar, J.: Computer-aided diagnosis of type 2 diabetes mellitus using thermogram of open mouth. In: Progress in Advanced Computing and Intelligent Engineering, pp. 167–177. Springer (2018)

22. Stojanovic M, Cvetanovic G, A.A.M.S.D.R.N.: Risk factors in dental caries. Cent Eur J Public Health 26(2), 104–110 (2018)

23. Thiele, S., Peltner, J., Richter, A., Mensink, G.B.: Food purchase patterns: empirical identification and analysis of their association with diet quality, socio-economic factors, and attitudes. Nutrition journal 16(1), 69 (2017)

24. Trevino, V., Falciani, F.: Galgo: an r package for multivariate variable selection using genetic algorithms. Bioinformatics 22(9), 1154–1156 (2006)

25. Twigg, S.M., Wong, J.: The imperative to prevent diabetes complications: a broadening spectrum and an increasing burden despite improved outcomes. The Medical Journal of Australia 202(6), 300–304 (2015)

26. Winkleby, M.A., Jatulis, D.E., Frank, E., Fortmann, S.P.: Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. American journal of public health 82(6), 816–820 (1992)

27. Zanella-Calzada, L.A., Galván-Tejada, C.E., Chávez-Lamas, N.M., Galván-Tejada, J.I., Celaya-Padilla, J.M.: Multivariate features selection from demographic and dietary descriptors as caries risk determinants in oral health diagnosis: Data from nhanes 2013–2014. In: Electronics, Communications and Computers (CONIELECOMP), 2018 International Conference on. pp. 217–224. IEEE (2018)